THE LIMITS OF SCREENING

Joseph Wu

King's College,

Cambridge

August 2019

This thesis is submitted for the degree of
Doctor of Philosophy.

DECLARATION OF ORIGINALITY

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

STATEMENT OF LENGTH

This dissertation, excluding the bibliography, contains 67,274 words. It does not exceed the word limit of 80,000 words set by the Degree Committee of the Department of History and Philosophy of Science, University of Cambridge.

THE LIMITS OF SCREENING

Joseph Wu

This thesis is about screening for cancer—about testing healthy individuals for disease. The traditional way to approach this subject is to begin by considering whether screening is effective. So, for example, there is a hearty debate about whether cancer screening reduces mortality to a meaningful degree. Some people claim that it does; others claim that it does not. But all seem to think that determining whether screening reduces mortality will resolve the controversy around screening. The motivating idea behind this thesis is that focusing on effectiveness obscures a deeper and different set of issues. The central claim is that understanding and justifying cancer screening requires attention to its moral dimensions. Ethical issues arise throughout the debate—not only in constructing a theory of effectiveness, but also in deciding evidentiary thresholds, in adjudicating which principles should guide screening policy, and in judging how to cope with risk and uncertainty. The subsequent chapters aim to show how significant progress and clarity can be achieved in the screening debate with some aid from ethics. Taken together, these chapters develop a framework for thinking about how cancer screening can and should be justified.

ACKNOWLEDGMENTS

I am indebted to many people for this thesis. Thanks to Tim Lewens, my supervisor, for the remarkably insightful guidance and encouragement throughout the PhD. His kindness is rivalled only by his ability to bring out the best in people.

Thanks to Stephen John, formally my advisor but informally my co-supervisor, for serving as an academic compass as I navigated my research. His intellectual influence on this thesis should be apparent to anyone who knows Steve or his work.

My interest in philosophy of science was sparked at Duke. Thanks to Rick Gawne, who nurtured my philosophical inclinations when I was confused but curious. His mentorship throughout the years has been invaluable. To Alex Rosenberg, for instilling in me the importance of making bold arguments. To Dan McShea and Robert Brandon, whose outstanding lectures in the philosophy of biology opened my eyes to an entirely different way to see science. And to Ray Barfield, who taught me that medicine not only relies on philosophy, but requires it.

For comments or discussion of ideas in this dissertation, thanks to Louise Barr, Azita Chellappoo, George Nikolakoudis, Lynette Reid, Raphael Scholl, Reuben Shiels, Michael Thornton, William Wong, and Ron Zimmern.

I am eternally grateful to Gates Cambridge, not only for funding this research, but also for providing a community of brilliant people to remind me of the myriad ways that research can improve the world. Over the years, I have also met outstanding people from the Cambridge Swimming and Water Polo Teams, the City of Cambridge Swimming and Water Polo Club, Rainbow Rocket Climbing Gym, King's College, and Darwin College. I am deeply grateful for all of the support I have received through these communities.

At last, thanks to my parents and my sister, Jean, for providing a tremendous amount of love and encouragement over the years.

TABLE OF CONTENTS

Introduction **1**

Chapter 1: Screening **8**

1.1 Introduction 8
1.2 What is Screening? 11
1.3 The Logic of Screening 15
1.4 Wilson and Jungner's Screening Criteria 22
1.5 The Current Screening Debate 26
    1.5.1 The Evaluative Issues
    1.5.2 The Epistemic Issues
1.6 Moving Forward 42

Chapter 2: Non-Maleficence **43**

2.1 Introduction 43
2.2 The Do No Harm Dilemma 45
2.3 Three Objections 51
2.4 *Ex Ante* Do No Harm 56
2.5 The *Ex Ante* Pareto Principle 61
2.6 Calculating Prospects 68
2.7 Conclusion 74

Chapter 3: Effectiveness **75**

3.1 Introduction 75
3.2 Mixed Claims 77
3.3 Two Views of Effectiveness 80
    3.3.1 Aggregate Population Effectiveness
    3.3.2 *Ex Ante* Individual Effectiveness
3.4 Effectiveness as Value-Laden 92
3.5 Pluralism about Effectiveness 97
3.6 Conclusion 105

Chapter 4: Uncertainty **106**

4.1 Introduction 106
4.2 The Risk-Generalization Model 109
4.3 The Mechanistic Model 117
4.4 Taking Stock 123
4.5 Implications of Uncertainty 125
    4.5.1 Prospects
    4.5.2 Non-Maleficence
    4.5.3 Effectiveness
    4.5.4 Reasonable Evidence
4.6 Conclusion 140

Chapter 5: Underdetermination **141**

5.1 Introduction 141
5.2 Evidence and Values 145
    5.2.1 Underdetermination
    5.2.2 Inductive Risk
5.3 An Objection: Betz on Value-Freedom 150
5.4 The Value-Apt Ideal 156
5.5 Mammography Wars 161
5.6 Conclusion 163

Conclusion **165**

Appendix I: Measurement Challenges **171**

Appendix II: Base Rates and Screening **175**

Bibliography **177**

THE LIMITS OF SCREENING

Introduction

This thesis is about cancer screening. It is about the practice of offering tests for cancer to otherwise healthy individuals. One might find this topic odd, however, for this is a thesis in the *philosophy* of science. Cancer screening seems to be a topic firmly rooted in the disciplines of public health, or clinical medicine, or biomedical science. This is true. There has been extensive discussion in all of these fields around different aspects of screening, from how to improve the accuracy of screening tests to how to measure the effectiveness of screening. But sifting through this large medical literature, one finds that screening occupies a curiously ambivalent position in the minds of health professionals. Indeed, the issue of cancer screening is one of the most hotly debated practices in the field (Bleyer and Welch 2012; Esserman, Shieh, and Thompson 2010; Gigerenzer 2015; Gøtzsche 2009; Shieh *et al.* 2016)

On the one hand, the underlying motivation for screening is straightforward and intuitive: early-stage cancer is easier to treat than late-stage disease. So, the reasoning goes, if we can detect cancer earlier, then we can improve cancer-related health outcomes. It seems to follow that we should offer screening for the earlier-than-otherwise detection of cancer. On the other hand, many health professionals are deeply skeptical that screening is the right way to go about improving cancer-related health outcomes (Gøtzsche 2012; Welch and Brawley 2018). They point out that the evidence underwriting the claim that 'screening saves lives' is either biased or suggests a very small reduction in mortality, at best. And they point out that, viewed from this more pessimistic perspective, screening boils down to unnecessarily testing millions of people to save a rather small number of lives.

These issues cluster around a general problem, which is the overarching question of the thesis: "Should cancer screening programmes be offered?" The focus of this thesis is predominantly on

the conceptual and ethical dimensions of this question. I did not undertake the gargantuan task of conducting a randomized trial, nor did I perform any qualitative interviews about, say, the lived experience of screening. Again, this may strike some as peculiar. Anecdotally, when I mention to peers that I work on cancer screening, many are confused about how the methods of philosophy can contribute to what might appear to be a predominantly scientific or medical issue. How, they inquire with an air of politeness that hardly disguises their suspicion, can my presence in libraries, perched in front of my laptop, generate the hard evidence to adjudicate debates about screening?

It is certainly true that empirical disputes lie at the heart of whether screening programmes should be offered, as they do for many questions in science and medicine. But conceptual and ethical clarity are prerequisites, too. Unless we can agree on what it means for a screening programme to be effective, the question of whether screening should be offered has little prospect of empirical resolution, no matter how much data is collected. What is more, as this thesis strives to make clear, I am sceptical that there is a sharp distinction between empirical questions, on the one hand, and conceptual and ethical questions, on the other. This is supported, for instance, by the recent philosophical literature on the proper role for ethical and political values in scientific inquiry, which I examine more fully in Chapter 5 (Biddle 2016; Douglas 2000; 2009; John 2014; 2018; Lewens 2019; Steele 2012). Focusing on the conceptual and ethical questions, then, is not intended to downplay the significance of empirical data, but rather to help clarify how such evidence relates back to our overarching question concerning whether screening should be offered at all.

In recent years, philosophers of science have begun to work on issues related to cancer biology. A conference I attended in Bordeaux in 2018, for example, was dedicated solely to the 'Philosophy of Cancer Biology.'[1] Moreover, there has been some insightful work, in the past few years, analyzing cancer through the lens of philosophy of science. Anya Plutynski (2018), in a recently published book, examines the conceptual and methodological issues that arise in cancer research, unpacking how scientists in molecular biology, epidemiology, and evolutionary biology strive to understand the disease. Lucie Laplane (2016), in another recently published book, draws on philosophy to clarify the fierce disagreement, amongst scientists, over the role of stem cells in

[1] https://www.philinbiomed.org/event/philosophy-of-cancer-biology-workshop/

cancer initiation and progression. And Marta Bertolaso (2016), in yet another recently published book, examines the underlying ontological commitments of different cancer research programmes, drawing out how explanatory models of cancer rely on distinct assumptions about, say, the nature of causation. All of these works are fascinating in their own right, yet despite a shared focus on cancer, this thesis will attempt to carve out a different path that does not fit neatly into this recent work on the philosophy of cancer. For one thing, my interest here in particular is on *screening* for cancer, a topic which has received scant attention from philosophers—barring one or two exceptions (Reid 2018; Rogers *et al.* 2017). And for another thing, the arguments that follow will not limit the focus to only metaphysical and epistemological issues, as all of the above works do; rather, a substantial portion of the forthcoming arguments will underscore the normative dimensions of cancer screening. Indeed, one of the chief aims of this thesis is to develop and apply some tools from moral philosophy to advance the screening debate.

Two additional comments to clarify the strategy of this thesis will be helpful. The first concerns how I view the role of philosophy in relation to the screening debate. I do not think that philosophical issues are merely "secondary" to scientific or medical ones. Rather, philosophy has a far greater role to play. Philosophy can (and should) serve a "complementary" role to science and make fruitful contributions to scientific and medical disputes (Chang 2004).

Here is one way of giving this admittedly abstract point some appeal: while reading the literature on cancer screening, I have continually been struck by how many positions about screening make implicit philosophical assumptions at key junctures in the argument. Sometimes, these assumptions are relatively innocuous. Equating the "effectiveness" of a screening programme with "improvement of length or quality of life" seems broadly reasonable. But other assumptions are far more controversial. Equating the "effectiveness" of a screening programme with the "ethical permissibility" of that same programme is a moral mistake, as I argue in Chapters 2 and 3. It is a vitally important task to be aware of these assumptions, to scrutinize and make them explicit, and to reflect on just how much our thinking about screening hinges on these subtleties. This is a task that falls naturally to philosophers of science and medicine. Illuminating these assumptions is another major aim of this thesis.

The second comment concerns the scope of this thesis. A well-worn strategy in projects of this length is to focus primarily on the conceptual issues, or primarily the epistemic, or primarily the ethical issues of a given topic. For example, the recent works in the burgeoning field of the "philosophy of cancer," discussed above, exemplify this tailored focus. By comparison, this thesis comprises a chimera of topics; it aspires to examine screening through all of the aforementioned lenses. Is this varied approach a weakness of the thesis? Certainly, it means that I have sacrificed some depth, say, in abstaining from going very deeply into the epistemological foundations of medicine. But I take some consolation in the fact that, in recent years, there has been some remarkably insightful work in the epistemology of medicine (Stegenga 2018; Howick 2011; Broadbent 2013). What is more, I think this trade-off in depth is worthwhile, for it yields the advantage of enabling me to explore a wider breadth of issues in the thesis. This is important because screening is such a multifaceted health practice, with elements of clinical medicine, and public health, and epidemiological research, and the psychology of decision-theory, amongst many others. In light of this complexity, it seems more interesting and fruitful not to limit our approach to just one philosophical angle. Instead, it seems to me that burrowing deeply enough just to achieve the necessary clarity from a given perspective, before stepping back and taking a more comprehensive viewpoint, will yield the most noteworthy pay-offs—both philosophically and practically.

This last point requires clarification. James Wilson (2009) makes a crucial distinction between two aims of normative theorizing about public health practices, like screening. One aim is *practical*. Sometimes, ethical theorizing is undertaken to usefully guide public health policy. Another aim is *conceptual*. Other times, ethical theorizing is undertaken to illuminate a more abstract set of issues, for instance, about the nature of "effectiveness." This distinction is important, because as Wilson (2009) notes, making progress on the *practical* set of issues does not automatically entail progress on the *conceptual* set of issues, and *vice versa*. This thesis has both of these aims in mind. Chapter 2 will develop and defend a normative principle to guide screening policy. The goal will be to alter how policymakers think about the ethical permissibility of a screening programme, with direct upshots for practice. By contrast, Chapter 3 will explore the implications of these arguments for a conceptual question about "effectiveness." Chapters 4 and 5 will occupy a middle ground between these *practical* and *conceptual* aims. Chapter 4 begins with a *conceptual* aim, clarifying the

relationship between a debate in political philosophy about how to cope with uncertainty and a debate in the scientific community concerning the ambiguous evidence around the harms of screening. It then turns to the *practical* aim, drawing out the implications of that relationship for screening policy. Likewise, Chapter 5 begins with a *conceptual* aim, examining the relationship between a debate in philosophy of science concerning inductive risk and the formulation of screening policy, before concluding with some comments that speak to the *practical* aim of how to implement a good screening programme.

Accordingly, while this thesis is written primarily for philosophers of science and medicine, it was my intention for the ideas developed here to be illuminating and useful for physicians, public health policymakers, and any parties working on the early detection of cancer. After all, my overarching question is one that is of interest to many across the health professions. And the subsequent chapters, taken together, are intended to provide a general framework for thinking through different dimensions of cancer screening. What I hope to do here is to reframe our thinking about cancer screening, motivated by a sensitivity to ethical issues. These ethical issues have been underexplored and unresolved in discussions of screening. Thus, if there is one central claim that this thesis aims to advance, it is that making progress in the screening debate requires careful reflection on these underlying ethical nuances. There are *moral* reasons, in addition to broadly *medical* reasons, that should influence what we think and do about screening.

Here is a more detailed synopsis of the main ideas and arguments advanced in each chapter:

In Chapter 1, "Screening," the key concepts central to understanding screening are explained and the logic of screening is spelled out. I then introduce the Wilson and Jungner screening criteria, which forms the basis of standard approaches to evaluating screening, and explain how these approaches are ethically wanting. On the one hand, standard approaches presuppose a range of salient ethical issues that warrant scrutiny. On the other hand, these approaches are silent on how these issues should be adjudicated. These considerations motivate the need for an ethical framework to guide the formulation of screening policy. The chapter concludes by discussing the current debate over screening. A distinction between "evaluative issues" (What are the benefits and harms of screening? How should these be balanced?) and "epistemic issues" (What is the

evidence for these benefits and harms of screening?) is introduced. The distinction is helpful in clarifying the key junctures of the screening debate. The "evaluative" and "epistemic" issues in the screening debate are then explained.

In Chapter 2, "Non-Maleficence," a tension between the principle of non-maleficence and screening programmes is introduced. This raises a dilemma: either all screening programmes are unethical or the principle of non-maleficence should be rejected. Given that neither option is appealing, the chapter argues that the best way to dehorn the dilemma is with a novel interpretation of non-maleficence in *ex ante* terms. This underwrites a principle, *ex ante* DNH, with direct upshots for screening policymaking. In particular, *ex ante* DNH implies that any screening programme which lowers the prospects of some invited is *prima facie* impermissible. The relationship between this principle and two significant debates in political philosophy and philosophy of science is discussed. The debate in political philosophy concerns the proper "currency" of distributive ethics; the debate in philosophy of science concerns how to calculate "prospects."[2]

In Chapter 3, "Effectiveness," I argue that the concept of screening effectiveness is far more value-laden than commonly presumed. In service of this argument, I articulate two different ways of thinking about screening effectiveness. After sketching the attractiveness of both interpretations, I show that they each rely on different implicit moral stances. This value-ladedarness is not problematic *per se*, but it leads naturally to a pluralistic view of effectiveness. The chapter concludes by spelling out how these arguments relate to other accounts of effectiveness.

In Chapter 4, "Uncertainty," the arguments developed in Chapters 2 and 3 are extended to contexts of uncertainty, where decision-makers are not in a position to assign precise probabilities to outcomes. In the first half of the chapter, I discuss Fuller and Flores' (2015) Risk Generalization-Particularization Model, which they dub the "standard model of prediction in medicine." I argue that the assumptions required to relate population frequencies to individual prospects often do not hold in the case of overdiagnosis. This underwrites an argument against the view that decision-

[2] This chapter draws on material written for an unpublished journal manuscript on non-maleficence and screening. Thanks are due to my co-author on that paper, Stephen John.

makers can assign precise probabilities to an individual's chances of overdiagnosis. After addressing an objection appealing to mechanistic models of prediction, the second half of the chapter develops the implications of these imprecise probability assignments for a theory of prospects, the principle of non-maleficence discussed in Chapter 2, and the concept of effectiveness discussed in Chapter 3.

In Chapter 5, "Underdetermination," the philosophical literature concerning the role of non-epistemic values in science is discussed. The "argument from inductive risk" is explained, and the relationship between this argument and the problem of evidentiary thresholds in the context of screening is explored. I argue that non-epistemic values can and should influence the setting of evidentiary thresholds in cancer screening. Moreover, I suggest that there are reasons to require a high evidentiary threshold, before implementing a new screening programme, due to issues around how screening programmes, themselves, alter societal and individual values.

I hope that these arguments can shed light on some interesting problems in cancer screening. Of course, one might point out that I have not provided a definitive answer to the motivating question of the thesis: "Should cancer screening programmes be offered?" After all, I do not defend a definitive *yes* or a definitive *no*, instead settling for the ever-frustrating response: "it depends." That is of no help, one might say. We wanted guidance from philosophy, and we got none. But I respectfully disagree. What this thesis aims to do is to identify precisely where "it depends," to develop the contours of the issues we need to resolve in order to make progress in the screening debate. The approach I am exploring here may not leave the reader completely satiated. It may not definitively resolve the screening stalemate. However, rather than think this is for a lack of benefit from philosophy, I think this speaks more to why screening is so stimulating and ripe for philosophical analysis, as we will see.

CHAPTER 1

Screening

1.1 Introduction

In January 2019, the National Health Service (NHS) unveiled its Long-Term Plan, outlining how it will translate a £20 billion budget increase into "better care for major health conditions." One of the key elements of this plan focused on cancer care. But a closer inspection reveals that the emphasis is not merely on any type of cancer care. The NHS will not invest in the research and development of better drugs for treatment, nor will they invest in training more oncologists or cancer nurses or cancer scientists. Rather, the plan exhibits a nearly exclusive focus on cancer care of a specific form: the *early detection* of cancer. The rationale is simple. The plan explains: "one of the biggest actions the NHS can take to improve cancer survival is to diagnose earlier." And the plan sets out an ambitious milestone: "By 2028, the NHS will diagnose 75% of cancers at stage 1 or 2." Roughly speaking, achieving this will mean that, "from 2028, 55,000 more people each year will survive their cancer for at least five years after diagnosis."[3]

Detecting cancer earlier is a very appealing thought. Increasing survival for 55,000 people, each year, is no trivial affair. Just imagine you, or someone you love dearly, is one of those people who will benefit. But sometimes very appealing thoughts need to be inspected closely, if only because they seem too good to be true. Matters are rarely so simple. Here is one complication: only about half of all cancers, at present, are diagnosed at stages 1 or 2. This raises the question of how, exactly, the NHS will increase early detection to three quarters of cancer diagnoses.[4] Conveniently,

[3] https://www.longtermplan.nhs.uk/

[4] A very different complication arises in interpreting what '75% of cancers' even means. One might mean 75% of all cancer *types* diagnosed early. But this seems implausible. Some cancer types just do not avail themselves to the strategy of early detection. Screening for ovarian cancer often requires *surgery* to confirm a diagnosis (Jacobs 2016). Undergoing the scalpel hardly seems worthwhile just to find out whether you *might* have abnormal pathology. Alternatively, one might mean 75% of all cancer *diagnoses* fall in Stage 1 or 2. This is more plausible, but it faces

there is data on the "main routes to cancer diagnosis." Around 34% of cancers are diagnosed via an urgent general practitioner (GP) referral with a suspicion of cancer, around 25% by a routine GP referral, about 20% from Emergency Presentation, around 11% from inpatient or outpatient appointments, and only around 6% are screen-detected.[5]

I mention these statistics for a reason. They raise an implication that needs to be taken seriously. As it turns out, the vast majority of cancer diagnoses occur in the clinic, via referrals from the GP or from presentation in A&E. So, if the goal is to diagnose many more cancers early, then either people will (implausibly) need to start noticing their symptoms earlier or the number of asymptomatic people tested for cancer will need to dramatically increase. The Long-Term Plan understandably goes the latter route, setting out the aims of lowering the starting age for bowel cancer screening, implementing more cervical cancer screening tests, overhauling breast cancer screening to increase uptake, and deploying vans in supermarket car parks equipped with mobile lung scanners.

The zealous optimism for the early detection of cancer reflected in the Long Term Plan, however, belies a contentious debate over the wisdom of testing asymptomatic individuals for disease. In the case of breast cancer screening, for instance, some have vocally argued that the harms from breast cancer screening outweigh any benefits (Baum 2013). One variant of their worry is this: sometimes, screening detects harmless cancers. And usually, these harmless screen-detected cancers are treated. But cancer treatments are not risk-free. Radiotherapy, for instance, increases one's risk of heart disease and lung cancer (Darby *et al.* 2013). It follows that as more people are treated with radiotherapy, the more people will die from heart disease and lung cancer. And it follows that as more *harmless* cancers are treated with radiotherapy, there will be a certain tipping point—a brink at which these lives lost from the toxicity of radiotherapy will offset the lives saved from screening. Some critics of screening claim that current screening practice is beyond this precipice, that the breast cancer deaths avoided by screening are offset (and maybe even outweighed) by other deaths caused by screening (ibid.).

---

its own set of difficulties. These are elaborated on in Section 1.3. In brief: we cannot equivocate 'early detection of cancer' with 'better cancer outcomes.' Sometimes 'early detection' is just plain harmful.

[5] http://www.ncin.org.uk/publications/routes_to_diagnosis

Others have raised a more modest worry. These skeptics of screening do not explicitly denounce screening for having an unfavorable benefit-harm ratio, but instead point out that, in all the trials conducted thus far, no screening programme has demonstrated a statistically significant decrease in overall mortality (Prasad *et al.* 2016). Here is one variant of this worry: it is very common to hear that "screening saves lives." But to take this claim at face value would be a mistake. The evidence on screening suggests that screening saves lives *from breast cancer*; it does not suggest that screening saves lives *overall*. So, by analogy, a drug that reduces your risk of a fatal heart attack may save lives from heart disease, but if it also leads to lethal strokes it may not save lives overall. What matters, ultimately, is whether a drug saves lives overall. And the same is true with screening — what we should really be concerned about is whether our best available evidence suggests that screening reduces the overall number of lives lost. Our best available evidence does not suggest this (Saquib *et al.* 2015). So, we should harbor a decent amount of suspicion toward whether offering cancer screening is a good idea.

How exactly to parse and adjudicate these debates is a substantive issue that will preoccupy us in the following chapters. But one point is clear from the outset. Not only is screening already an important health practice, affecting millions of individuals each year, it will only become *more so* in the following decade. Many more people, in the near future, will be offered screening—they will be presented with the momentous decision of whether to get tested for a disease which, at that point in time, they do not believe they have. Given that screening will play an increasingly prominent role in healthcare practice, a framework specifically attuned to the philosophical problems raised by screening is worthwhile to develop.

This chapter sets the stage for the examination of screening that follows. Section 1.2 explains in more detail what screening is, and how it differs from clinical medicine. Section 1.3 fleshes out the logic of screening, and Section 1.4 introduces the standard criterion used to evaluate screening. Section 1.5 distinguishes between the "evaluative" issues and the "epistemic" issues underlying screening, before discussing what these issues involve.

1.2 What is Screening?

The NHS defines screening as follows: "a way of identifying apparently healthy people who may have an increased risk of a particular condition."[6] To get a grip on how this is a unique type of healthcare, consider the following cases:

> Clinical Medicine
> While chopping vegetables for dinner, after an exhausting day at work, you notice a lump on your right forearm. That is odd, you think, but dismiss it as inconsequential. You reassure yourself that miscellaneous bumps are common from time to time. Over the next few months, the lump seems to grow in size and shift in color, becoming a firm, red nodule. It becomes painful when pressure is applied. You begin to get worried, so you schedule an appointment with a dermatologist. After examining the results of your skin biopsy, your physician provides a diagnosis: it is squamous cell carcinoma, a form of skin cancer.
>
> Screening
> While chopping vegetables for dinner, after an exhausting day at work, your partner says that you have a letter from the NHS that seems important. Surprised and confused, you clean the knife and use it to slice open the envelope. The letter is offering you the choice to get tested for bowel cancer. And, conveniently, the letter already includes an appointment at the local health center for your test. You read the leaflet, outlining the benefits and harms of screening, and after careful reflection decide to attend. And thank goodness you do! The screening leads to a diagnosis of bowel cancer, in its earliest stage. Your prognosis is good.

Both scenarios are broadly concerned with the diagnosis and treatment of cancer. But there are (at least) three important differences between the cases. The first is that in Clinical Medicine, you have symptoms. You notice that something is off, and suspect that the nodule may be a health problem. By contrast, in Screening, you have no symptoms, and believe yourself to be healthy with respect to the disease for which you have been offered testing. The second is that in Clinical Medicine, the decision to schedule the appointment with a health professional originates from yourself. You could have ignored the nodule for longer, or you could have scheduled an appointment at first sight of the lump. The point is that you were the one to contact, and by

[6] https://www.nhs.uk/conditions/nhs-screening/

extension initiate a relationship with, your physician. By contrast, in Screening, the test for bowel cancer is offered to you. Absent the letter, you would not have been tested for bowel cancer. It was your healthcare system that reached out to you. The third is that in Clinical Medicine, the focus of the medical intervention is *you*—the one with symptoms, who sought out the appointment. By contrast, in Screening, the target is strictly speaking a certain *population*, of which you are a member. As Raffle and Gray (2007) explain in their classic text on screening, screening functions as a 'sieve' or 'filter' at the population-level, the goal of which is to identify select individuals whom might require further investigation.

Here is one way to visualize the process:

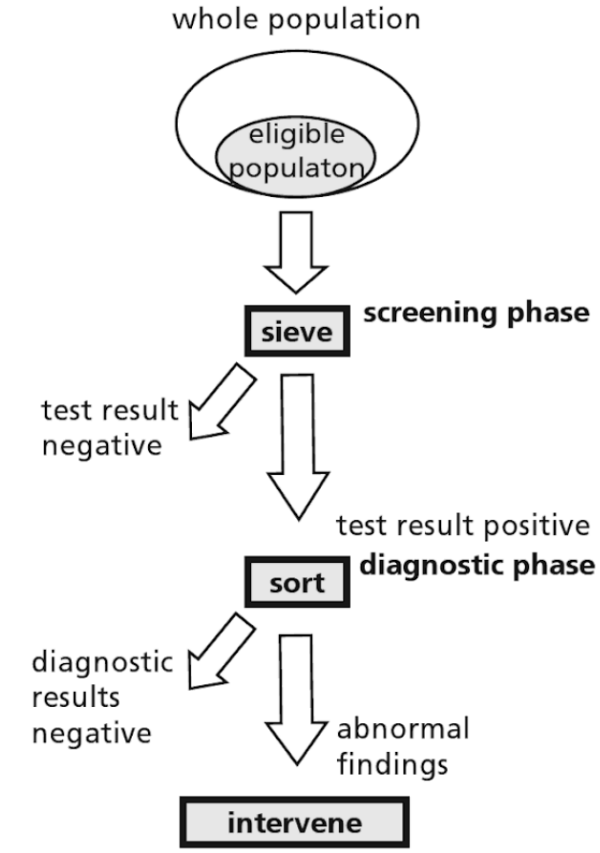

Figure 1. The basic steps of a screening programme. A certain population, deemed to be at increased risk, is offered a test which 'sieves' out those who might require further investigation. In turn, individuals are then 'sorted' between those with disease and those without disease (Raffle and Gray 2007).

Admittedly, Figure 1 is an oversimplification. Sometimes test results are equivocal, and not all abnormal findings lead to intervention. Moreover, different screening programmes may not exactly fit the pathway above. With chlamydia testing, for example, there is no separate sorting phase and intervention is determined on the basis of the initial test. But Figure 1 is nonetheless illustrative, broadly speaking, of how a screening programme operates. Moreover, our focus is specifically on cancer screening programmes. For our purposes, Figure 1 is representative of the types of programmes we will be examining in this thesis.

So screening is different from standard clinical medicine in that it is offered to asymptomatic individuals in a population deemed to be at increased risk of, say, cancer. At this point, two distinctions are worth making to clarify the scope of the subsequent arguments. My focus in this thesis will be on screening *programmes*, which involve inviting large numbers of people to get tested. These are generally understood as a public health service, organized by the state, that attempts to 'filter' out individuals whom might require medical intervention (Wright and Zimmern 2014).[7] Examples in the NHS include the breast cancer screening programme, which triennially invites women between 50-70 years of age; the bowel cancer screening programme, which invites people for a one-off bowel scope screening test at age 55 before then inviting men and women between 60 to 74 to do a stool test every two years; and the cervical cancer screening programme, which triennially invites women between the ages of 25 and 49 and invites women between the ages of 50-64 every 5 years.

Of course, within the "category" of screening programmes, there can be broad differences in how screening is carried out. In the United States, various institutions publish guidelines which recommend individuals to get screened, beginning at a certain age, and at certain intervals. This system of recommending screening is very different from NHS screening programmes, and these differences may give rise to different ethical issues. So, for example, "what" it is permissible to claim about screening may be ethically sensitive to "how" that claim is made: it is one thing for me to *offer you the choice* to invest all your life savings in my risky, innovative start-up, and quite

[7] Some screening programmes have a separate aim: to provide information, even if the risk of adverse health cannot be changed by early intervention. An example of this is pre-natal genetic diagnosis. Since nobody tries to justify cancer screening as a way to provide information, as opposed to preventing future ill health, I set these issues aside.

another thing for me to *recommend you* to invest all your life savings in my risky, innovative start-up. Chapter 5 explores these issues in more detail. The essential point, for now, is that the focus of this thesis will be on cancer screening programmes within the NHS in particular.

But screening programmes are not the only way in which screening can be carried out. Suppose you go to a medical appointment for back pain and your physician, noting your age and smoking habits, offers you a test for lung cancer. This *opportunistic* screening is different from a screening programme in how the test is offered. Whereas opportunistic screening involves being offered a test for an ailment unrelated to the issue that led you to the doctor's office, programmes involve tests being offered systematically to an entire group of individuals deemed to be at increased risk of disease. Opportunistic screening raises interesting ethical issues in its own right, but it is not my focus. Nonetheless, many of the central ethical ideas developed in the subsequent chapters can straightforwardly be generalized to opportunistic screening.

The second distinction concerns primary and secondary prevention. Primary prevention concerns the *prevention* of disease, for example, by identifying and removing the precursors of cancer. Pap smear tests for cervical cancer screening fit this bill. This screening modality aims to identify preinvasive cervical intraepithelial neoplasia which, if removed, reduces the incidence of cervical cancer. Alternatively, secondary prevention concerns the *early detection* of disease. Mammography for breast cancer, prostate-specific antigen (PSA) testing for prostate cancer, chest X-rays and CT scans for lung cancer, and faecal occult blood test (FOBT) fit this bill. These modalities strive to identify existing disease at an earlier-than-otherwise stage. Accordingly, these modalities do not reduce the incidence of disease but try to reduce morbidity and mortality. The divide here is, however, not sharp. Colonoscopy can both prevent and detect bowel cancer earlier-than-otherwise; pap smears can both prevent and detect cervical cancer earlier-than-otherwise. And whether a given test prevents or detects disease turns, in part, on when abnormal cells count as "genuine disease." But this is a tricky matter, because drawing the line between disease and predisease is not straightforward (Schwartz 2014). We will return to these issues. For now, let me just stress that I will not be concerned much with the primary and secondary prevention distinction. I will abstract from this distinction, focusing instead on the benefits and harms that stem from a given screening modality. This is, I take it, what matters most for an account of the ethics of

screening programmes.

1.3 The Logic of Screening

Proposals for widespread screening of asymptomatic disease can be traced back more than a century. Screening, for instance, is often connected with the advent of the periodic health examination (Han 1997; Croswell, Ransohoff, and Kramer 2010). In his 1861 treatise *Lectures on the Germs and Vestiges of Disease, and on the Prevention of the Invasion and Fatality of Disease by Periodical Examinations*, the British physician Horace Dobell proposed that the routine application of physical examinations and laboratory tests on asymptomatic individuals could be used to discover the "earliest evasive period of defect in the physiological state" (Dobell 1861). Dobell's conviction was that periodic health exams were "the only means by which to reach the evil and to obtain the good" (ibid.). Meanwhile, in the United States, a similar movement toward testing symptomless individuals was brewing. The physician George Gould, at the fifty-first annual meeting of the American Medical Association, proclaimed: "He is the greatest discoverer who finds the presymptom…There are a thousand undiscovered…advance scouts and forerunners, to be learned when the slight and unconscious departures from normality are studied by examinations of the supposedly well" (Gould 1900).

Soon these ideas expanded to influence the approach to cancer. In *The Control of a Scourge, Or How Cancer is Curable*, published in 1907, Dr. Charles Childe lamented that the loss of life from cancer is avoidable:

> Early cancer has no symptoms....The victims...are quite naturally lulled by the entire absence of symptoms into a sense of security....The ignorance of the importance of the early sign is so great, the temptation to make light of the apparently trivial symptom so natural....[But] cancer itself is not incurable...it is the delay that makes it so....If every case of cancer came under the notice of the [physician] at the earliest possible moment...it requires no stretch of the imagination...to say that the majority...would be cured (Childe 1907).

Dr. Childe framed the high mortality rate of cancer as a problem of late diagnosis. And this focus on early disease detection became increasingly cemented in medical practice in the twentieth

century. In the 1920s, the American Medical Association suggested that the periodic health exam for early disease detection was so obviously beneficial that medical evidence (which, during this period, consisted of anecdotal case reports) was not necessary for its implementation: "Medical experience of the benefits of periodic examinations of presumably healthy persons is sufficiently widespread to make any detailed reference superfluous" (Emerson 1923). Screening for multiple conditions gained further popularity, after World War II, with the finding that many young military recruits had conditions which disqualified them from service (Reiser 1978). This general interest in screening dovetailed with an increasingly specific interest in early cancer detection (Mukherjee 2010). There was an upsurge of large-scale health campaigns to "fight" the "war" against cancer, highlighting the importance of being vigilant for early cancer warning signs. A frequent slogan was "Delay Kills!" (Lerner 2001). Other slogans minced their words even less. A campaign poster from the American Cancer Society in the 1980s declared that: "If you haven't had a mammogram, you need more than your breasts examined" (Gigerenzer 2014).

These historical and political forces shaping early disease detection are fascinating in their own right, but I set them aside henceforth. The discussion above is useful, however, because it underscores a guiding principle underlying the logic of cancer screening:

> *Early Detection Principle* (EDP): the earlier a condition is found, the better the chances of a positive health outcome (Harris *et al.* 2011)

So, for example:

> Abdominal Pain
> Mike is a 59-year-old engineer. He loves to cycle and hike on the weekends, and often indulges his affinity for Italian cuisine in the evenings. Recently, he has been experiencing abdominal pain and, after eating, he has been bloating an unusual amount. Initially, Mike dismisses these as the result of stress. But soon the symptoms begin to affect his ability to cycle, hike, and consume linguine carbonara. After his physician runs some tests, Mike learns that he has bowel cancer. The prognosis is not good. Stage IV.
>
> Bowel Screening
> Three years earlier from Abdominal Pain, Mike the 56-year-old engineer has just returned from a long hike along the beach, during which he watched the sun set in

> the foggy distance. In his mail, he finds a letter from the NHS, offering a bowel scope screening test through the national bowel cancer screening programme. Mike decides to attend. During his appointment, his physician finds and removes some polyps from his bowel, and sends them to the lab for testing. Mike soon learns that he has bowel cancer. But the prognosis is good. Stage I.

Many people may be familiar with situations similar to Abdominal Pain. In these scenarios, a common tendency may be to think: "This is truly devastating. Had the cancer been detected earlier, then I may have had a better chance at surviving longer." And many people may be familiar with situations similar to Bowel Screening. In these scenarios, the general tendency may be to think: "Thank goodness I attended screening! Had the cancer been detected far later, then I may have had far worse chances for survival." Indeed, broadly speaking at least, the EDP appears to be true for many cancers. There is a strong association between health outcome, understood as survival, and the stage at diagnosis. For example: 98% of people diagnosed with Stage I bowel cancer survive for at least one year; only 40% of those diagnosed at Stage IV survive for at least one year. 100% of people diagnosed with Stage I breast cancer survive for at least one year; only 63% diagnosed at Stage IV survive for at least one year. And 99% of those diagnosed with Stage I cervical cancer survive for at least one year; only 35% diagnosed at Stage IV survive for at least one year.[8]

We need to be extremely careful with our statistics here. It tempting to look at the stage-specific survival rates above, to note that survival is much better at earlier versus later stages of cancers, and to immediately conclude that the EDP is true for (say) breast cancer. If you found yourself making this inference, worry not—you are not alone. It is common for journal articles, media reports, and advocacy groups to use survival rates to justify screening. One recent academic article argued that, based on increased survival rates, the higher costs for cancer medicine in the United States versus Europe are "worth it" (Philipson *et al.* 2012). Similarly, the Komen Foundation, in its breast cancer awareness campaign, repeatedly cites higher breast cancer survival for screened women as a central consideration for why women should undergo mammography (Woloshin *et al.* 2012).

[8] See: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/survival; https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/survival#heading-Three; https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/cervical-cancer/survival#heading-Three

But I said that the EDP only "appears to be true" above for a reason. There are complications with survival rates. And these complications are especially important to bear in mind when dealing with early detection. Survival rates refer to the proportion of cancer patients still alive, at a given time, after diagnosis. Consider:

> Clinical Survival Time
> Mary notices an oddly shaped lump on her left breast. She is clinically diagnosed with Stage III breast cancer at age 53. Tragically, she passes away from breast cancer two years later, at the age of 55.

Survival rates "start the clock" at the time of diagnosis. So in this case Mary's survival time, from diagnosis, is two years. She was diagnosed at age 53 and passed away at age 55. But there is a problem. Sometimes screening can artificially inflate survival statistics. Here is one way: screening may appear to improve survival merely because you start the clock sooner, by detecting cancer earlier, but with no difference in overall outcomes. Consider:

> Lead Time Bias
> Three years earlier from Clinical Survival Time, Mary, at the age of 50, receives her first letter from the NHS inviting her to breast cancer screening. After some thoughtful deliberation, Mary accepts the offer. Mammography detects a small clump of cellular abnormalities. Stage I breast cancer. Mary begins treatment. Tragically, she passes from breast cancer five years later, at the age of 55.

In this case Mary's survival time, from diagnosis, is five years. She was diagnosed at age 50 and passed away at age 55. The time Mary survives with breast cancer is higher in Lead Time Bias than in Clinical Survival Time, an increase from two years to five years. But to think this increase in survival time represents a genuine medical benefit—to think that it improves length or quality of life—seems dubious. Screening made no difference in clinical outcomes, for in both scenarios Mary passes away at age 55 from breast cancer. So the benefits suggested by the increase in survival time between the cases above are illusory, the result of adding "lead in" time onto Mary's recorded survival. Lead time bias is not the only way in which survival rates can be misleading. Screening may appear to improve survival because it preferentially detects slower-growing cancers

with good prognoses (length time bias). Or screening may appear to improve survival because it detects cancers that will never progress to cause harm (overdiagnosis bias). These biases are explained in more detail in Appendix 1. But the point is that survival rates can improve even when no deaths are delayed or prevented. The point is that we need to be wary about the phrase "positive health outcome" in the EDP—for screening to be beneficial, it is not enough for survival time to improve. Screening must also reduce mortality or improve quality of life compared with no screening (Cho *et al.* 2014).

Survival statistics settled, what kind of principle is the EDP in relation to screening policy? For starters, it is surely not a *sufficient* condition to implement screening. There is a gap between the claims "the EDP is true for this cancer" and "screening should be implemented for this cancer." Setting aside some niceties, the easiest way to see this is to note that "the EDP is true for this cancer" is consistent with screening imposing serious harms on many people. This can occur if, for instance, the disease one is screening for is very rare, and if as a result the number of false-positives is very high.

You might find this point about many false-positive results odd. Most screening modalities are rather good at correctly identifying the presence or absence of disease (in more formal jargon, the tests have high sensitivities and specificities, respectively). But even when this is the case, a curious paradox can sometimes arise, when the rates of false positive tests become unsettlingly high. This "false positive paradox" arises because the baseline incidence of disease is sometimes very low. So, for example, imagine Cambridge has two cab companies, Yellow and Blue. Yellow cabs make up 95% of the cab population, blue cabs 5%. Imagine that yesterday, just as dusk was settling, a cab struck an unlucky postgraduate, right in front of King's College, before racing off into the distance. Suppose we know eye witnesses are 95% accurate at correctly discerning blue cabs (in more formal jargon, eye witnesses have a sensitivity and specificity of 95% for correctly identifying blue cabs). Suppose a passerby claims that a blue cab committed the hit-and-run. What is the chance that the cab was actually blue?

It is tempting to answer 95%. After all, this is how accurate eye witnesses are! But this would be mistake. To claim this would be to ignore information relevant to the situation—the base rate of

cabs. The correct answer, somewhat surprisingly, is 50%. A more detailed explanation and proof is provided in Appendix 2. But the takeaway is straightforward enough. Even if eye witnesses are rather good at correctly identifying blue cabs, the chances they are accurate diminishes when the underlying incidence of blue cabs is low. The same reasoning generalizes to screening for disease. If the underlying incidence of cancer in a screened population is very low, as is often the case, then the number of false-positive results may be rather high. What is more, there is evidence that false-positive cancer screening results can lead to long-term psychosocial harm (Brodersen and Siersma 2013). If a screening programme causes more net harm than net benefit for a population, then this seems to be good grounds for thinking that the programme is unacceptable.

Returning to the EDP, the underlying issue here is that the EDP is concerned solely with the potential benefits of screening. But, as will be continually emphasized throughout this thesis, it is a costly moral mistake to view screening solely through the lens of benefits. This threatens to obscure and overlook the harms of screening, explained more fully below, which has been a recurrent tendency in the past.

Perhaps, then, the EDP is a *necessary* condition for screening policy. This seems more plausible. At the very least, for screening to be justified, there must be a benefit stemming from earlier versus later detection of cancer (Harris *et al.* 2011). Even so, it would be a mistake to presume that the EDP *is* always satisfied with screening. Consider:

> Good Later-Stage Treatment
> The Pap Smear reveals a low-grade cervical intraepithelial neoplasia (CIN) in Jane. Her prognosis is good. Breathing a sigh of relief, Jane counts her blessings and vows to appreciate life more. But in the counterfactual in which Jane skipped her appointment, and was not diagnosed with CIN until symptoms arose, her prognosis would have been roughly equivalent. Even if she were diagnosed with high-grade CIN, there are good treatments to excise all of the abnormal cells in her cervix.
>
> Bad Early-Stage Treatment
> During a CT scan to examine a broken bone, a radiologist incidentally discovers that Sarah has early-stage pancreatic cancer. Her prognosis is bad. There are no treatments available that can prolong her survival. Sighing with grief, Sarah resigns herself to enjoying and appreciating every experience she has left in this world. In

> this scenario, early detection may actually make Sarah worse off than later detection, for she is simply aware, for a longer period time, that there is little that can be done to alter her fate.

In both scenarios above, there is no benefit to earlier detection (even though, in Bad Early-Stage Treatment, Sarah fares better in terms of survival time compared with later diagnosis!). What both scenarios above bring out, and what is especially worth highlighting is this: the EDP rests on assumptions about the availability and effectiveness of treatments in a given context. If there is no benefit from earlier versus later *treatment*, then this in turn implies that there is no point in earlier versus later *detection.* This is important, because it underscores that screening is not merely "a way of identifying apparently healthy people who may have an increased risk of a particular condition."[9] Screening also aims to benefit these individuals via treatment. Cancer screening programmes are *inter alia* treatment programmes.

It is worth noting that the EDP, as formulated above, limits the justification of screening programmes to "increased chances of positive health outcomes," rather than "positive consequences more generally." For example, imagine two programmes that equally increase the chance of positive health outcomes for those invited, but one is far cheaper for the NHS to implement. Of course, these considerations of "efficiency" are also relevant to justifying screening. Given the limited resources of the NHS, it is plausible to think there is reason to implement the cheaper programme rather than the more expensive one. The EDP does not claim that the *only* justification for a screening programme can be positive health outcomes. However, it does seem plausible that the EDP is the *primary* justification for a programme. This is supported by the way in which the NHS communicates about screening. For example, the leaflet inviting women to breast cancer screening explains, on its first page, that "The NHS offers screening to save lives from breast cancer." While it may be possible to justify cancer screening on other grounds, this would need to be made adequately transparent. In current screening practice, which is the focus of this thesis, the main justification stems from positive health outcomes.

[9] https://www.nhs.uk/conditions/nhs-screening/

To summarize, the EDP is a necessary condition to justify screening, but it is not sufficient.[10] And whether the EDP is true, in a given context, cannot simply be presumed true. Sometimes it is just plain false, such as when there is rather good late-stage treatment or rather bad early-stage treatment. What other conditions need to be satisfied to get to good screening policy?

1.4 The Wilson and Jungner Criteria

When the interest in screening began to heighten in the early twentieth century, what drove the appeal of early detection was, in large part, the development of new tests. And the underlying reasoning was just so intuitive (Croswell *et al*. 2011). Of course more testing is a good idea, just like checking your email is a good idea. However, not much consideration was given to how screening affects health outcomes—to whether checking your inbox five times a day makes any positive difference to just checking once a day. Things began to change in the 1960s and 1970s. Some began to raise questions about the wisdom of introducing unexamined screening programmes. The science of clinical epidemiology was being developed, and "evidence-based" groups started to raise challenging questions. This more critical approach to screening led to the development of more comprehensive guidelines for the evaluation of screening programmes (Harris *et al*. 2011).

The most influential guideline was published, in 1968, by the World Health Organization. Authored by Wilson and Jungner, these 10 "principles" continue to guide screening policy in countries where a programme is being considered, and have formed the basis of all other screening criteria proposed since the monograph was published. The principles are as follows (Wilson and Jungner 1968):

[10] You might worry whether the EDP, strictly speaking, is a necessary condition. Stephen John raises the following possibility: imagine that treatment is equally effective at all stages of cancer. It follows that the EDP is false. And imagine that treatment is much cheaper at earlier versus later stages of cancer. It follows that we may have efficiency-related reasons to still implement screening. So, the EDP may not be a necessary condition for screening. Fair enough. I am perfectly happy to concede this point, in this hypothetical scenario. But in actual practice, this worry will not hold water. Cancer is much more difficult to treat at later stages, after it has invaded and metastasized other tissues. This will be true for the foreseeable future, and so it is not the case that treatment is or will be equally effective at all stages of cancer. Our focus is on actual practice. Given this focus, it is plausible to understand the EDP as a necessary condition.

Wilson and Jungner Criteria

1. The condition sought should be an important health problem
2. There should be an accepted treatment for patients with recognized disease
3. Facilities for diagnosis and treatment should be available
4. There should be a recognizable latent or early symptomatic stage
5. There should be a suitable test or examination
6. The test should be acceptable to the population
7. The natural history of the condition, including development from latent to declared disease, should be adequately understood
8. There should be an agreed policy on whom to treat as patients
9. The cost of case-finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on medical care as a whole
10. Case-finding should be a continuing process and not a "once and for all" project

These criteria remain the starting point for discussions around screening today. And they help to flesh out, in more detail, the logic of screening. For example, the second criterion—that there should be an accepted treatment for patients with recognized disease—accounts for one worry raised above about the EDP. As we saw, in the absence of good early-stage treatment when disease is detected earlier-than-otherwise, screening fails to confer any benefit. Hence, the second criterion blocks the issue raised by Bad Early-Stage Treatment. What is more, the first criterion—that the condition sought should be an important health problem—mitigates some of the worries around the harms of screening. For example, Croswell *et al.* (2010) interpret the notion of "an important health problem" in terms of the disease burden in the population and the overall risk-benefit ratio of utilizing mass screening for a given group. The latter issue may be affected by the prevalence of disease because, as we saw above, the prevalence of disease impacts the rate of false-positive test results (Appendix 2). So the first criterion can be understood as one way of helping to assure that a screening programme leads to more overall benefit than harm for a population.

Despite its popularity and influence, there are reasons for doubting that the Wilson and Jungner criteria offer a fully comprehensive account of how to evaluate screening. As Juth and Munthe (2012) note, the principles themselves provide no guidance on how they should be interpreted. For our purposes, it is worth highlighting how the Wilson and Jungner criteria presuppose substantive

ethical issues whilst simultaneously providing no guidance on how to settle them. For example, consider Criterion 7: "The natural history of the condition, including development from latent to declared disease, should be adequately understood." But just when is a condition "adequately understood"? With our privilege of hindsight, we now know that population screening was initiated at a time when our understanding of cancer was dominated by a linear model of cancer progression. On this picture, cancer was thought to progress along a series of increasingly injurious steps, with each progressive step invariably following the former. The thought with screening, then, was that early detection of these cellular abnormalities would "break the chain" of disease development and ultimately benefit the individual.

However, it turns out that some cancers may not progress to cause harm at all. Take prostate cancer, a relatively slow-growing disease that affects primarily older men (Welch and Albertsen 2009). Many elderly men have prostate cancer, but their disease is often indolent, causing little or no harm. Hence, a platitude often invoked is that most men die *with* prostate cancer, but not *of* prostate cancer.[11] And it turns out that some cancer may even *regress*. Take cervical intraepithelial neoplasia (CIN), a precursor of cervical cancer often detected by screening. CIN is classified into four grades. It turns out that the majority of CIN I lesions will spontaneously regress within a year, and recent evidence suggests that about half of CIN II lesions will, too (Castle *et al.* 2009). Since treatment for CIN can have negative effects on fertility and pregnancies, intervening on low-grade CIN that would have otherwise regressed results in net harm for that individual (Sadler *et al.* 2004). Presuming that all cancers follow a linear model of progression, then, led to many people being unnecessarily treated for benign lesions. This was a costly error.

You might wonder why this matters. Of course it was an unfortunate error, but mistakes in medicine are not uncommon. How, exactly, does this relate to when a condition is "adequately understood" for the purposes of screening? The crucial link comes from an influential "argument from inductive risk" in philosophy of science. This inductive risk argument is a substantive topic I will return to in Chapter 5, but a very brief introduction will serve to illustrate the point here. According to this argument, the moral consequences of error should indirectly impact where we

[11] In 2018, at an early detection of cancer workshop, I attended a talk by a urologist whom pointed out that the highest rates for prostate cancer are in the 90+ age group, according to Cancer Research UK. And, the urologist coyly noted, this is well past the median life-expectancy for UK males.

set the evidentiary threshold for accepting a scientific claim (Douglas 2009). Contrast, for example, how much evidence should be required to accept that a certain drug is safe with how much evidence should be required to accept that a batch of belt buckles is not defective. Plausibly, we should require more evidence to accept that the drug is safe, because "how sure we need to be before we accept a hypothesis will depend on how serious a mistake would be" (Rudner 1953, 2).

Here is how the argument from inductive risk might run for CIN: a claim such as "cervical cancer follows a linear model of progression" is a scientific claim. Since whether we should accept or reject this claim is not deductively entailed by the available evidence, the question then becomes how much evidence is sufficient for accepting that "cervical cancer follows a linear model." But deciding whether to accept this claim requires a value judgment about the tolerability of different sorts of errors. Specifically, that value judgment depends on how bad it would be to err, either in mistakenly accepting the (false) claim or failing to accept the (true) claim. The argument, very crudely, might run as follows: how bad would it be to fail to implement screening, if cervical cancer does follows a linear model? And how bad would it be to implement screening, if cervical cancer does not follow a linear model? Deciding the seriousness of these mistakes is usually understood to reflect the "moral standards" of those making the value judgment (ibid.).

So the argument from inductive risk, if correct, suggests that far from being straightforwardly interpreted, Wilson and Jungner's Criterion 7 actually presupposes a rather substantive set of ethical issues concerning when we have "adequately understood" cancer progression for the purposes of screening policy. And the Wilson and Jungner criteria, taken alone, do not provide any guidance in this normative domain. The principles are silent on whether ethical value judgments should impinge on evidentiary thresholds *at all*, as well as on the further issue of if they should have an influence, which values are permissible or desirable.

This is not a drawback *per se* of the Wilson and Jungner criteria.[12] The principles were never formulated with the explicit intention of addressing the ethical complexities of screening. But I take the comments above to illustrate two important points, which will be continually emphasized in this thesis. The first is that the ethics of screening have been vastly underappreciated and

[12] Nor of any other "standard" approach to screening that has been influenced by the Wilson and Jungner Criterion (Harris *et al.* 2011). There are many different "guidelines" for screening policy, but none take a stance on the ethics of screening.

underexplored. The "argument from inductive risk" is merely one of many points at which ethical value judgments intersect with our thinking about screening. As Chapter 2 argues, a key ethical principle that has been overlooked in discussions of screening is the principle of non-maleficence. The second point is that fleshing out the logic of screening warrants a closer scrutiny of just how ethical issues underpin the controversy around screening. Determining whether the criteria for screening are met, somewhat surprisingly, requires taking a stance on moral issues.

1.5 The Current Screening Debate

Practicing good screening on a population-level has been very controversial. Some of these controversies are relatively straightforward. When thousands of women were erroneously not invited to cervical cancer screening in 2018, there was uproar over the mistake.[13] Similarly, when thousands of women were erroneously not invited to breast cancer screening, also in 2018, there was uproar over the mistake.[14] These controversies are straightforward, in one respect, because they stem from a failure of the healthcare system to deliver expected care. Imagine your physician has atrocious bedside manner. There is a clear sense in which you might object, in which you might be angry and annoyed at her sheer incompetence, precisely because she is not fulfilling her duties of care.

Other controversies are trickier. In the breast cancer screening error, for example, once the initial uproar subsided, there was an interesting twist. A letter to the *Times* co-signed by 15 medical professionals intimated that the breast cancer screening error was, as a matter of fact, a "lucky escape": "The breast screening programme mostly causes more unintended harm than good, has no impact on all cause mortality, and claims of lives 'saved' are counteracted by deaths resulting from interventions" (Hawkes 2018). And this view is not exactly far-fetched; it is not merely a vocal minority opining their skepticism about breast cancer screening. These medical professionals have good company with some very esteemed clinical researchers (Gøtzsche 2012, Welch 2006, Baum 2013).

[13] https://www.bbc.co.uk/news/health-46208137
[14] https://www.bbc.co.uk/news/health-43973652

This controversy concerning whether the breast screening error was a "lucky escape" is trickier because it conflates two very different types of issues. The issues have frequently been muddled in debates around screening, but it is important to keep them distinct:

> *Evaluative issues*: What are the actual benefits and harms of screening? Taken together, do they constitute a favorable benefit-harm ratio?
>
> *Epistemic issues*: What are we licensed to conclude, given the evidence, about the benefits and harms of screening? When is the evidence sufficient for acting on the claim that screening is effective?

It is often not clear which is the source of disagreement in screening debates. We can see this by holding one issue fixed and varying the other. Imagine, for example, we are absolutely certain that breast cancer screening saves one life for every 200 women tested, but leads to unnecessary treatment in three others. Maybe you think this is a favorable benefit-harm ratio; and maybe I do not. We disagree about the *evaluative* issues concerning what benefit-harm ratio must be before implementing screening. There is, to be sure, a wide range of views on these evaluative issues. Van den Bruel *et al.* (2015) conducted a population survey of attitudes toward unnecessary cancer detection and treatment, finding that some people (roughly 3-7%) are *not* willing to accept any level of unnecessary detection and treatment to save one life from cancer. Other people (roughly 7-14%) were willing to accept *a thousand people* being unnecessarily detected and treated for cancer to save one life from cancer. So there is, very clearly, considerable space for disagreement on evaluative issues.

Here is a very different way of disagreeing. Imagine that you and I both agree that a favorable benefit-harm ratio is one life saved for every 200 women tested at the cost of (no more than) three unnecessarily treated. Imagine also that you and I agree that screening should be implemented once we are 80% certain that this benefit-harm ratio would be achieved with the programme. However, we disagree about whether the available evidence translates into 80% certainty that this benefit-harm ratio would be achieved. Maybe you think we should go ahead with screening, justifying this thought with the trial evidence suggesting that screening reduces breast cancer mortality between 15-20%. Perhaps I am more cautious, basing my judgment on the trial evidence suggesting that no screening programme has reduced overall mortality. I say more on this below,

but the point is that, in this case, we disagree about the *epistemic* but not *evaluative* issues.

Debates around screening often involve disagreement about both issues. Arguably, this is to be expected. The issues are deeply interlinked, insofar as one's position on the evaluative issue may plausibly influence the epistemic issue. Here is an analogy: if the first study on a novel experimental therapy showed an overwhelmingly favorable benefit-harm ratio, then the drug may be sped to approval for clinical practice, even if there is still more epistemic uncertainty around the actual effect size than is usual to approve a drug. One way to understand this practice is in terms of the evaluative issue "shifting" the evidentiary standards for approval. Charting the terrain between the two issues is a complex task, to be explored in more detail in Chapters 3 and 5. The important point, for present purposes, is that the "evaluative" and the "epistemic" issues are separate, and keeping the distinction in mind helps to chart different sources of disagreement in screening.[15]

The remainder of the chapter is structured to reflect this divide. I first introduce the benefits and harms of screening. How does screening putatively help or harm those who are tested? Then, in the remainder of the section, I describe the evidence for the prevalence and magnitude of these benefits and harms. I focus on breast cancer screening as my case-study. I do this, in part, for rhetorical purposes: it is the screening programme that has evoked the fiercest controversy in the medical community. But I also do it, in part, for constructive purposes: precisely because breast cancer screening has been so controversial, it is the area that can benefit the most from careful philosophical analysis.

1.5.1 The Evaluative Issues

The Benefits of Screening

The two main benefits of cancer screening are the improvement of length and/or quality of life. As we saw above from Abdominal Pain and Bowel Screening, sometimes screening leads to the earlier-than-otherwise detection of cancer, and sometimes this means that the disease can be

[15] Thanks are due to an email exchange with Stephen John and Ron Zimmern for immense help in formulating these ideas.

treated at an earlier-than-otherwise stage. Treatments for early-stage cancer tend to be much better than late-stage cancer, in the sense that early-stage treatments reduce cancer incidence or mortality more than late-stage treatments. Screening can also, in some cases, improve quality of life. Imagine a woman who is deeply anxious that she has breast cancer, because a close friend was recently diagnosed. Imagine she is especially fearful of being diagnosed, too, because she witnessed firsthand the grueling nature of her friend's treatment regimen. And imagine the wave of relief she feels when the test comes back negative, confirming with high certainty that she is disease-free. Regardless of whether this woman's psychological beliefs are reasonable, screening confers a benefit by mitigating her anxiety. Even if screening does not actually prolong her life, it seems to be beneficial by improving quality of life.

The Harms of Screening

A curious feature of screening programmes is that most participants will not actually benefit from getting tested. Some harms, then, include the loss of time and financial resources (Smith *et al.* 2015). Millions of individuals will take time out of their busy lives each year, drag themselves to what can be an uncomfortable and invasive medical procedure, and ultimately derive little to no clinical benefit. Other potential harms include the medicalization of healthy individuals (Verweij 1999). Screening targets people whom do not believe they have cancer, but treats them as if they were patients. This expansion of the "medical" into "ordinary" aspects of human life can bring with it the unwelcome implication that there *may* be something wrong with you, and this in turn can increase anxiety about health (Aronowitz 2015). These are all very serious matters. However, my focus will be on three rather different harms of screening. These harms are the most troubling and acute ones that result from screening: false-positive test results, overdiagnosis, and unnecessary treatment. I elaborate on these in turn below.

False-Positives

Screening tests for cancer either come back positive, suggesting the presence of disease, or negative, suggesting its absence. However, many screening tests come back positive when, in fact, the individual does not actually have the disease in question. Typically, this means the individual must undergo further testing to confirm whether she has cancer. A tempting thought with false-positive results is that it is "better to be safe than sorry," and that once a false-positive finding is

confirmed to be benign one's psychological state will return to normal. Think about fire alarms. Every ear-shattering fire alarm I have experienced was (fortunately) a "false-positive." There was never an actual fire; just smoke from the popcorn some inebriated person forgot in the microwave. I may be cranky and irate to be jolted awake in the middle of the night, but I sure am happy the alarm went off, even though there was no fire. When I collapse back into bed, I doze off with the peace of mind that the building is not engulfed in flames. And I am no worse off from the false alarm. I harbor no new anxieties about suffering awful, fire-related injuries.

But the harms that result from false-positive screening results are of a different nature. They cannot be so easily dismissed. Brodersen and Siersma (2013) provide evidence that false-positive findings from mammography screening lead to long-term psychological harm. For a range of psychosocial outcomes such as anxiety, sleep-quality and sense of dejection, women with false-positive results consistently reported worse psychosocial states compared with women with normal findings. And of particular interest, these negative psychosocial states do not fade away entirely over time. Even up to three years after being declared free of potential cancer, women with false-positive results reported worse psychosocial outcomes compared with women with normal findings. So these women do not leave screening with peace of mind. They are, in an important sense, worse off from the false alarm.

Overdiagnosis

Overdiagnosis occurs "when a condition is diagnosed that would otherwise not go on to cause symptoms or death" (Welch and Black 2010, 605). Cancer overdiagnosis can be accounted for in two ways. On the one hand, the diagnosed cancer may never progress or, in some rare cases, even regress on its own. Recall the case of early-stage CIN from above, which usually resolves itself without intervention. Another example comes from ductal carcinoma in situ (DCIS), the abnormality most commonly detected by mammography. DCIS, however, only progresses to invasive disease 20-30% of the time, and many women can live with DCIS without ever experiencing any harm. The idea that some cancers are nonprogressive may seem mysterious, but a few physiological mechanisms that can halt the progression of cancer have been uncovered (Mooi and Peeper 2006). Some cancers might outstrip the available resources in their blood supply and, in consequence, starve; others might be recognized by the immune system and contained; still

others might simply not be aggressive at all to begin with. Chapter 4 explores the nature of cancer progression in more detail.

On the other hand, the cancer may be progressive, but develop slowly enough such that the patient dies of other causes before any symptoms occur. Recall the case of prostate cancer from above. We are now in a better position to understand just why screening for prostate cancer in very elderly males is, broadly speaking, a bad idea. Because small low-grade prostate cancer progresses so slowly, many men with prostate cancer never experience symptoms and, without screening, would never even know they have prostate cancer. Moreover, because many men over 75 have competing risks of mortality, there is not much point in screening for prostate cancer when individuals in this age group may die of other causes. The phrase "develop slowly enough" must be understood in a relative sense, though. Strictly, even cancers that are destined to be harmful can be overdiagnosed if they are detected when very small and in individuals with limited life expectancy.

Overdiagnosis is different from a false-positive test result. The former correctly identifies cellular abnormalities that fulfill the diagnostic criteria for cancer, whereas the latter incorrectly suggest the presence of cancer. Another way to see this is to refer back to Figure 1. False-positives occur at the "screening phase," whereby the "sieve" of screening fails to accurately filter out disease-free individuals. At the "diagnostic phase," however, it is confirmed these individuals are in fact disease-free. By contrast, overdiagnosis occurs solely at the "diagnostic phase." Overdiagnosed individuals do have "abnormal findings," because they do have cancer. It is just that the cancer is not and will not become harmful. It is worth noting, though, that the distinction here is a slippery one. The difference between false-positives and overdiagnosis appears to turn on "diagnostic categories." Yet, in the face of high rates of overdiagnosis, one might raise doubts that the diagnostic categories are, themselves, apt for clinical practice. So, for example, Schwartz (2014) argues that DCIS should not be categorized as cancer, but rather as a *risk factor* for cancer — like smoking for lung cancer. In these cases, there is something thorny about the relationship between concepts of disease, diagnostic categories, and overdiagnosis (Walker and Rogers 2017; Reid 2017). But aside from a brief discussion in Chapter 2, these very interesting and important issues are placed to one side in the thesis.

As Welch and Black (2010) put it, "the impact of overdiagnosis can be life-long and affects patients' sense of well-being, … their physical health, and even their life expectancy" (611). However, there is an interesting feature to overdiagnosis, which is that rarely does anyone know they have been overdiagnosed. We have good evidence that overdiagnosis is occurring in populations, as we will see in the following section. But it is nearly impossible to specify which particular individuals have been overdiagnosed, since "overdiagnosis can only be identified in an individual if that individual 1) is never treated and 2) goes on to die from some other cause" (Welch and Black 2010, 606). The trouble is that most screen-detected cancers are treated. So the way in which overdiagnosis harms people is strange—it is not like being punched in the face, after which you *know* you have been harmed in some way. Your bleeding, broken nose is evidence of that. The harms stemming from overdiagnosis are subtler. You may suffer the distress of being labeled a "cancer patient," the torment of facing your own mortality, but you are unlikely to experience any aggravation at the thought that screening made you worse off. On the contrary, you may genuinely believe that your cancer diagnosis is *why* you are still alive today.

It may be tempting, then, to dismiss the harmful nature of overdiagnosis. But this would be a serious mistake. The "invisibility" of harm at the individual-level is no reason to deny that overdiagnosis is a harm of screening, just as it is no reason to think that smoking is not harmful because we cannot pinpoint which individuals' lung cancer was specifically caused by cigarettes. These issues raise a number of conceptual and ethical considerations, which will be explored more deeply in Chapter 2. The main takeaway for now is that overdiagnosis is one of the most pernicious harms of screening. I should note, however, that there are two, slightly different ways in which overdiagnosis might be harmful. Overdiagnosis may itself be harmful, by unnecessarily turning people into patients. The very experience of being diagnosed with (harmless) cancer can be injurious (Willig 2011). Quite a different harm arises when overdiagnosis leads to overtreatment. The following section explains.

Overtreatment

Overtreatment occurs when therapy is "inappropriately invasive or extensive in relation to the biology of disease" (Shieh *et al.* 2016, 3). It is closely linked to overdiagnosis, insofar as overtreatment is generally a downstream consequence of overdiagnosis. By definition, intervening

on overdiagnosed cancers is medically unnecessary treatment. But there is some conceptual space between these notions. Not all overdiagnosed cancers are treated. For example, a treatment strategy for DCIS gaining traction is "watchful waiting," in which DCIS is closely monitored to see whether it will become life-threatening instead of immediately treated (Marshall 2014). Overtreatment can also occur even when a cancer is not overdiagnosed. Imagine a grueling and over-the-top chemotherapy regimen for a mildly harmful, medium-risk cancer.

Conceptual nuances aside, I do not think there is any point in mincing words here: cancer treatment is a horrific experience. Chemotherapy ravages the body; surgery disfigures it. It is a serious harm to put anyone through a cancer treatment regimen, especially if it was entirely unnecessary. The problem is not merely that overtreatment results in more harm than good; it is that such harmful interventions—a direct consequence of early detection programs—were never going to benefit the individual *at all*. In what follows I will sometimes use "medically unnecessary treatment" to refer to "overtreatment." The two notions can sometimes come apart (taking ibuprofen for minor headaches can be "overtreatment" but still "medically necessary" in some sense), but I will abstract from these subtleties.

1.5.2 The Epistemic Issues

So, screening is beneficial in that it can improve length or quality of life. Screening is also harmful, in that it can lead to false-positive results, overdiagnosis, and overtreatment. But it is one thing to know what the benefits and harms are, and quite another thing to know their prevalence and magnitude. The latter requires attending to what the evidence says, and to how that evidence relates to claims about the benefits and harms of screening. The present section discusses this latter set of "epistemic" issues. I will first discuss what the best available evidence suggests about the benefits of screening, and explain why getting a precise estimate of these benefits is a thorny issue. I will then discuss what the best available evidence suggests about the harms of screening. In Chapter 4, I will explain why getting a precise estimate of these harms is a difficult task.

Evidence of Benefit

For simplicity, let us focus on whether screening lengthens life, setting aside the impact of

screening on quality of life. As Appendix 1 explains, the basic problem with measuring the impact of screening is this: if all you do is measure health in screened people, then regardless of whether screening makes any relevant difference to length or quality of life, the conclusion reached will be one skewed heavily in favour of screening. In light of these worries, it is generally accepted that the best evidence for the benefits of screening comes from randomized controlled trials (RCTs).

The basic method of RCTs is this. A study population which is ideally representative of the target population is recruited, and randomly assigned to different groups.[16] One or more "study" groups receives the intervention of interest (i.e. they are screened triennially for cancer). Another "control" group receives standard care (i.e. they are not screened at all). The relevant outcomes (i.e. effects on mortality) are measured in the different groups, and compared. Here is the available evidence, from RCTs, examining the impact of mammography screening on mortality from a recent systematic review (Saquib *et al.* 2015):

> Four RCTs found a reduction in breast cancer mortality. In these four studies, pooled together, there were 633 screening deaths in a 170048 sample, and 746 control deaths in a 136889 sample. This results in a risk difference of 0.00173 or a number needed to screen of 578. However, it should be noted that the authors of the Cochrane Collaboration meta-analysis deemed these studies to have suboptimal randomization (Gøtzsche and Jørgensen 2013). Three RCTs with adequate randomization found no effect of mammography screening on breast cancer mortality. In these three studies, pooled together, there were 404 screening deaths in a 119504 sample, and 572 control deaths in a 172649 sample. This results in a risk difference of -.000067. No trials found a significant reduction in all-cause mortality.

A few issues with this evidence are worth noting.[17] First, there is the problem of how to amalgamate first-order evidence (what do the RCTs show?) and higher-order evidence (how reliable is this evidence?) (Sliwa and Horowitz 2015; Fuller 2018). Do we ignore the RCTs with suboptimal randomization, uncritically accept them, or accommodate them with a hefty pinch of

[16] Note, however, that in practice the study population is often not adequately representative of the target population. I explore this issue in more detail in Chapter 4.

[17] It should be noted that one of the co-authors of the Cochrane Collaboration study, Peter Gøtzsche, has been a controversial figure in debates about breast cancer screening, and was recently at the heart of a dispute at the Nordic Cochrane Collaboration in which he was expelled as Director (Burki 2018).

salt? So, for example, one way in which a trial was suboptimally randomized concerned the observation that more women in the control group than in the study group were diagnosed with breast cancer *before* entry to the trial (Gøtzsche and Jorgensen 2013). This implies that the trial results were biased in favor of screening, because the background rates of cancer are not evenly balanced between the "control" and "study" groups. But just how much, and in what ways, this bias impinges on what we should think of screening is a complex matter.

Second, there are well-known problems of transporting causal claims from one context to another. Roughly, good RCTs have high internal validity, so we can be reasonably confident that the effect size in the RCT is true — but only for the study population (Cartwright 2009). Transporting these results to the target population requires additional assumptions, some of which may not hold true (Fuller and Flores 2015). As Deaton and Cartwright (2018) emphasize, we cannot simply assume that the results of RCTs will export to new contexts. Rather, we need good reasons that can justify that the export is appropriate.[18] So, for example, the RCTs judged by the Cochrane Collaboration to have adequate randomization were conducted in Canada, Sweden, and the United Kingdom. If our interest is in modern day NHS breast cancer screening, we need good reasons for thinking that the results of these RCTs—even the one in the UK, given how much time has passed—will generalize to current practice. It may be that the different healthcare systems make a relevant difference, or it may be that the background rates of cancer in different contexts make a relevant difference. These issues will be explored more deeply in Chapter 4.

To say that the best evidence comes from RCTs, however, is not to say that evidence from "lower-quality" methods in the evidence-based medicine (EBM) hierarchy tell us nothing at all. After all, "which method is most likely to yield a good causal inference depends on what we are trying to discover as well as on what is already known" (Deaton and Cartwright 2018, 2). And, here is one thing we already know about screening: to lengthen life, screening must detect life-threatening disease at an earlier point, compared with no screening, in which it can be more successfully

[18] To do otherwise would be simple extrapolation, which faces well-known limitations. Consider Bertrand Russell's Chicken: a chicken infers, on all the available evidence, that the farmer comes in the morning to feed her. This inference is repeatedly confirmed. But then Christmas morning comes, and the farmer chops her head off and serves her for dinner (Russell 1912). The chicken's mistake was not her reasoning; it was that she did not understand the structural system within which she was existing and gathering her evidence.

treated.

This gives us an implication we can take to the historical data: if screening actually imparts benefits through earlier treatment, we can expect the incidence of early-stage (and more easily treatable) cancers to increase and the incidence of late-stage (and less easily treatable) cancers to decrease, following the introduction of screening programmes. Yet Bleyer and Welch (2012), in analyzing trends in breast cancer incidence the past three decades, find that there have been substantial increases in early-stage breast cancer incidence, but only a marginal reduction in advanced breast cancer incidence. This implies two things: first, that there is a good deal of overdiagnosis (for, otherwise, we would expect a concomitant decrease in late-stage incidence), and second, that screening is, at best, having a small effect on breast cancer mortality (for, otherwise, we would expect again a corresponding decrease in late-stage incidence). While the data was from the United States and hence raise worries about whether the conclusions transport to the NHS, it seems reasonable to think that Bleyer and Welch's analysis tells us something important about the impact of mammography on survival: we at least have reason to think that the true effect size will be on the lower end of our interval of plausible effect sizes (or, more cautiously, that the true effect size is unlikely to exceed the upper bound of our interval) (Bleyer and Welch 2012).

Here is something else know we already know: breast cancer mortality has been decreasing over time. This is a good thing. But it raises a question: to what extent is this due to better treatment for cancers independent of early detection, as opposed to the early detection of cancer via screening? Autier *et al.* (2011) conducted a trend analysis of breast cancer mortality in neighboring European countries with similar population structure, socioeconomic circumstances, healthcare services, and access to treatment. The interesting feature was that these matched countries also implemented breast cancer screening at different times. So, for example, mammography screening was implemented in Sweden in 1986, and implemented in Norway in 1996. If screening were effective at reducing breast cancer mortality, then we should expect a decrease in mortality in Sweden around 10 years earlier than Norway. But this is not observed. Though trends in breast cancer mortality decreased over time, the mortality rates varied little between Sweden and Norway. This implies that the reduction in breast cancer mortality over time is largely attributable to improvements in treatment irrespective of early detection, rather than screening for early disease

(Autier *et al.* 2011).

Evidence of False Positives

The NHS Breast Cancer Screening leaflet states that for every 100 women screened, four will have an abnormal test that requires further testing (four women are "sieved" out, as shown in Figure 1). But then of these four abnormal results, three will turn out to be a false-positive (in the "diagnostic phase" they are deemed to not have cancer, as shown in Figure 1).[19] Roughly two million are screened each year in the NHS Breast Screening Programme.[20] So when we scale up to actual practice, this number becomes much higher.

Why do so many women receive false-positive results? Recall from above the "false-positive paradox": even if test sensitivity and specificity are very high, there can still be a high number of false-positive results if the disease being tested for is rare (Appendix 2). Since screening affects healthy individuals as opposed to symptomatic patients, the baseline chance that a given woman has breast cancer is much lower than a woman in the clinic with, say, odd chest pains. And this means that the number of false-positives increases. Here is a stark example to demonstrate the point: we know that breast cancer risk increases with age. Most screening policies begin testing around age 50, and it is fiercely controversial whether, and how often, to screen women between 40 and 49 years of age. In large part, this is because the incidence of breast cancer for this age group is much lower, which entails a much higher false-positive rate. One estimate was this: "more than 1900 women would need to be invited for screening mammography in order to prevent just one death from breast cancer during 11 years of follow-up, at the direct cost of more than 20,000 visits for breast imaging and approximately 2000 false positive mammograms" (Quanstrum and Hayward 2009, 1076).

Evidence of Overdiagnosis

There are three forms of evidence to estimate rates of overdiagnosis. One, canvassed above, involves using historical data to estimate the extent of overdiagnosis. The hallmark indication of

[19] https://www.gov.uk/government/collections/population-screening-programmes-leaflets-and-how-to-order-them

[20] https://hansard.parliament.uk/Commons/2018-05-02/debates/BE9DB48A-C9FF-401B-AC54-FF53BC5BD83E/BreastCancerScreening

overdiagnosis is a rise in the incidence of early-stage disease alongside a minor or nonexistent decrease in late-stage disease. This is because the strategy of early detection is to "pull" advanced and more life-threatening cancers into an earlier, more treatable stage. As we noted already, at least for breast cancer screening, the historical data does not show this occurring.

Another source is autopsy studies. Many autopsies find disease reservoirs of cancers in individuals whom died of other causes. One study of American men and another of Greek men found disease reservoirs of prostate cancers between 30% and 70%. A study of thyroid cancer found disease reservoirs approaching 100%. A study of breast cancers in middle-aged women found ranges between 7% and 39% (Welch and Black 2010). The lifetime risk of harmful disease is much lower than these percentages. This implies that, when screening is implemented for (say) breast cancer, screening is uncovering a reservoir of disease that otherwise would not have resulted in clinical symptoms.

A third source, and the strongest form of evidence, comes from "catch up" studies. Here is the logic: we conduct a long-term follow-up randomized trial of screening. The screening group, we expect, will detect more cancers than the control group, merely because screening shifts the time of diagnosis earlier. If all of these cancers represent diseases that were destined to present in the clinic—that is, if none of these are instances of overdiagnosis—then this entails that the control group, over time, will "catch up" to the screening group in terms of the number of cancers detected. Why? Because all such cancers ought to present signs and symptoms. Thus, an excess of diagnoses in the screening group that persists long after the trial is very strong evidence that overdiagnosis has occurred. This is because, at least in theory, the only difference between the screening and control group is that the screening group underwent screening. Of course, in practice this may not be true, but these long-term follow-ups of RCTs are the strongest form of evidence that overdiagnosis has occurred because, given the nature of the randomized trial, the best explanation for the difference in the number of cancer diagnoses between the groups is overdiagnosis. While historical data and autopsies also provide good evidence for overdiagnosis, it is harder to ascertain whether the results of these observational studies are due to factors that are not overdiagnosis. Some estimates from "catch-up" studies: up to 24% of breast cancers picked up by mammography, 51% of lung cancers picked up by chest x-ray, and 67% of prostate cancers picked up by PSA

testing are cases of overdiagnosis (Welch and Black 2010).

However, estimates of the extent of overdiagnosis are highly variable, even when confined to breast cancer screening. A recent study found 16 estimates in the literature, ranging from 0% - 54% (Puliti *et al.* 2012). Different studies appeal to different bodies of evidence that, in turn, entail different estimates of breast cancer risk. Most studies examined the screening target population, not the women actually screened, so estimates were contingent on screening uptake. And critics of the higher estimates of overdiagnosis maintain that sufficient follow-up time was not granted. For example, some estimates only examined the first two or three rounds of screening. Yet there is reason to think that overdiagnosis occurs at a higher rate in these rounds compared to subsequent rounds. Hence, studying a larger screening period—say, 10 rounds over 20 years—may yield lower estimates. A more detailed analysis of these issues will be given in Chapter 4.

Evidence of Overtreatment

When individuals receive a cancer diagnosis, it is not exactly surprising that they will want to do something about it. After all, there is a common tendency to equate "cancer" with "mortality." Accordingly, Esserman *et al.* (2010) argue that low-risk cancers should be renamed *IDLEs*—indolent lesions of epithelial origin—to mitigate patient miscomprehension of the health situation. This proposal is motivated by the observation that, generally speaking, individuals with screen-detected cancers will opt to pursue treatment. The extent to which overtreatment occurs, of course, depends on the extent to which overdiagnosis occurs. Accordingly, getting a precise idea of just how often screening leads to unnecessary treatment is also a difficult task. But it is helpful to know that most NHS screen-detected cancers are, in fact, treated (Marmot *et al.* 2013). Roughly four fifths of women with screen-detected breast cancers receive radiotherapy (Baum 2013).

Outcome Measures

The studies above provide a broad strokes picture of the magnitude of benefits and harms in screening. At this point, you might wonder what all the fuss over screening is about. Imagine you are faced with the decision of whether to be screened. You know there is some chance it will benefit you, and you know there is some chance the test will harm you. So what? That just means screening, like most medical interventions, is *risky*. There is uncertainty about which outcomes

will eventuate, depending on what you choose. But this hardly seems different from the uncertainty surrounding whether a heart transplant operation will be successful. And there is far less controversy around whether heart transplants are a good idea, broadly speaking, for specific people. So why is screening so much more contentious than other medical interventions?

One key difference is the sheer amount of uncertainty around the magnitude of benefits and harms in screening. For example, in the independent review of the NHS breast cancer screening programme, the panel wrote: "overdiagnosed cancers certainly occur, but the frequency in a screening programme of 20 years duration is unknown" (Marmot *et al.* 2013, 2206). It is one thing for you to not know what will happen if you undergo heart surgery, and quite another thing for you to not have a sense of the chances the surgery will go smoothly or wildly awry. And it turns out to be surprisingly difficult to fine-tune our picture, to attain a very precise estimate of the benefits and harms of screening. Introducing the debate between disease-specific mortality (DSM) and overall mortality (OM) outcome measures is illustrative here.

A striking feature of the evidence about cancer screening is that no trial has demonstrated a reduction in overall mortality (OM). No trial has demonstrated that screening reduces deaths from *all* causes. There is disagreement about what follows from this observation. Some say this: "So what? Breast cancer screening reduces deaths from *breast cancer*; cervical cancer screening reduces deaths from *cervical cancer.*" Their thought is that screening was never intended to reduce OM, just as life-saving chemotherapy for metastatic cancer is not expected to reduce OM. What is more, demonstrating a reduction in OM is nearly impossible practically speaking. Because breast cancer contributes little to total mortality, a trial needs millions of participants in order to demonstrate a statistically significant reduction (Prasad *et al.* 2016). Hence, if we want to get a sense of the benefits of screening, we should look to the disease-specific mortality (DSM) measures.

Others, however, demur. They say DSM is often biased in favor of screening. For one, DSM ignores the fact that screening causes harm, which can offset the benefits in screening. By analogy, a drug that reduces deaths from stroke may reduce stroke mortality, but if it leads to a greater number of heart attack deaths it is clearly not beneficial all things considered. And for another,

DSM is sensitive to idiosyncratic biases in the study methodology. For instance, should DSM include deaths from a perforation due to screening colonoscopy, when that person did not have colorectal cancer? Such judgment calls can bias the results. And these concerns are avoided by looking at OM measures, which count the harms and are not biased with respect to how deaths are classified, hence being "a more reliable end point" (Penston 2011). What is more, it is worth pointing out that a recent meta-analysis of mammography screening trials suggested that only four screening programmes, out of seven considered, showed a significant reduction in DSM (Saquib *et al.* 2015). Hence, even if DSM were the only outcome measure used, it is far from clear that screening does, in fact, constitute a favorable benefit-harm ratio.

Stepping back from the screening debate, some commentators note that one reason OM outcome measures are advantageous is that they can capture unanticipated effects of interventions (Newman 2010). Some historical case-studies are illustrative. OM measures were used to discover the ineffectiveness, and in the worst cases the severe harms, of previously accepted therapies: fibrates for cholesterol reduction, antiarrhythmics, and red cell transfusion for the critically ill. Conversely, OM measures have been used to illustrate the benefit of treatments previously thought to be harmful, as was the case with beta-blockers for patients with chronic congestive heart failure. Why were these therapies erroneously accepted? There is a common culprit: the use of plausible surrogate outcome measures that, unbeknownst at the time, had other effects (Newman 2010). Though DSM used to be an accepted outcome measure for screening, critics of screening are beginning to re-cast it as a surrogate outcome measure for OM (Prasad *et al.* 2016).

When faced with no reduction in OM, one can go two ways on this. Some people argue that this demonstrates that any mortality benefits of screening are being offset by deaths from treatment. For example, Baum (2013) has argued that if we include the deaths caused by overdiagnosis and treatment, then the harms from breast cancer screening outweigh the benefits. Roughly, for every 10,000 women invited to screening, 3–4 breast cancer deaths are avoided, but at the cost of 2.72–9.25 deaths from the long-term toxicity of radiotherapy. Others argue that because trials with enough power to detect a difference in overall mortality are practically unfeasible, and because DSM suggests a small mortality benefit, there is some "net benefit" to screening, even though the

benefit is not so large as to show in OM measures (Steele and Brewster 2011). Regardless of which response one prefers, one point is clear from the outset: there is substantial epistemic opacity around the magnitude of the benefits and harms of screening.

1.6 Moving Forward

In this chapter I have endeavored to set the groundwork for the remainder of the thesis. I have tried to explain, in just the detail necessary for the forthcoming philosophical arguments, what cancer screening is, why it is interesting and important, and the logic underpinning screening. I drew a distinction between evaluative and epistemic issues, two often conflated junctures of disagreement. In so doing, I have tried to sketch the contours of how screening can be beneficial, how screening can be harmful, and how we can get a sense, based on the evidence, of the extent to which screening is beneficial or harmful. This was a long-winded introduction into screening, with not much philosophy. The subsequent four chapters will remedy this philosophical lacuna. In the following chapter, I restrict my focus to the evaluative issues, namely, the harms of screening. I will argue that these harms raise a tension with the principle of non-maleficence that needs to be taken seriously.

CHAPTER 2

Non-Maleficence

2.1 Introduction

In the previous chapter, we saw how screening is a highly controversial health practice. We saw how screening can greatly benefit some people. It can save lives through early detection. We also saw how screening can greatly harm some other people. It can lead to psychologically damaging false-positive results, overdiagnosis, and unnecessary medical treatment. How can we make some headway out of this debate? My starting point is to address the following question: "When is a screening programme ethically justified?" The aim of the present chapter is to develop and defend a relevant ethical principle to guide screening policy.

A tempting place to start is with the traditional bioethical principles guiding clinical practice (Beauchamp and Childress 2013):

> Beneficence: a physician is under a *prima facie* obligation to act for the medical benefit of her patients
> Respect for Autonomy: a physician is under a *prima facie* obligation to respect the decision-making capacities of autonomous persons
> Non-Maleficence: a physician is under a *prima facie* obligation not to cause medically unnecessary harm to her patients
> Justice: a physician and/or healthcare system is under a *prima facie* obligation to distribute benefits and harms fairly and equitably

These principles are firmly rooted in our understanding of ethical medicine. In their seminal work, Beauchamp and Childress (2013) advocate the principles as four *prima facie* standards that form a framework for ethical reasoning. As they note, however, the principles are abstract and divorced from practical contexts. To be useful, they need to be specified and balanced against each other.

Our task at hand, then, is two-fold. We need to decide which ethical principles should guide our thinking about screening, as well as how such principles should be interpreted and balanced against others. In the literature on the ethics of screening, most discussions focus on beneficence and respect for autonomy. For example, one way of understanding the (sometimes unwarranted) faith in early detection is in terms of an implicit appeal to beneficence. That same focus on beneficence seems to be reflected in the NHS Long Term Plan's near-exclusive focus on early detection as the strategy to improve cancer care. These considerations suggest that beneficence is already a guiding principle of screening policy. Equally, many have raised concerns about whether consent in screening is adequately informed (Gotzsche 2009; Gigerenzer 2014). And the official NHS strategy for communicating about screening places heavy emphasis on providing the opportunity for informed choice: the screening leaflets seek "neither to encourage screening nor to ask people to make decisions without guidance" (Forbes *et al.* 2014, 195). These considerations suggest that respect for autonomy is already a guiding principle of screening policy.

Of course, beneficence and respect for autonomy are important ethical principles in screening practice. These principles *should* guide screening policy. But there seems to be a curious lacuna in ethical reasoning about screening. While many commentators point out that screening imposes harms, that these harms may sometimes outweigh the benefits (Baum 2013), discussions of screening have had little to say about whether these harms generate ethical reasons to think that screening programmes are impermissible, or whether screening programmes might violate the non-maleficence principle. This is concerning, as it is analogous to moral reasoning in controversial ethical cases without awareness that respect for autonomy is a principle with moral valence. Our reasoning in such cases would be seriously deficient without at least considering how respect for autonomy can generate reasons to choose one course-of-action over another. Likewise, our ethical reasoning about screening is seriously deficient without considering how non-maleficence can generate reasons to formulate screening policy in one way or another.

The present chapter aims to shed light on the ethical contours of screening by emphasizing and sharpening the non-maleficence principle for screening policy. The starting point of the argument is that the imposition of screening-related harms is *prima facie* incompatible with the non-maleficence principle. From this seemingly innocuous observation, developed over the course of

the following two sections, we are led to a novel interpretation of the non-maleficence principle in *ex ante* terms (Section 2.4), to some surprising implications about the ethical permissibility of population health policies (Section 2.5), and to some insights about the ethics of risk (Section 2.6). I begin by describing the tension between screening and non-maleficence in the next section.

Another clarification may be helpful before proceeding. You might wonder about considerations of justice. I have emphasized the importance of autonomy, beneficence, and non-maleficence for screening, but have said little thus far about justice. For now, all I ask is some patience from the reader. Sections 2.5 and 2.6 of this chapter will discuss some intersections between justice and screening. As we shall see, one potent objection to the central argument for non-maleficence developed here will derive from an egalitarian principle concerning the fairness of outcomes.

2.2 The Do No Harm Dilemma

Consider the following example:

> Individual Doctor
> There are four women in Dr. Jones's exam room: Anne, who has life-threatening lung cancer, and Betty, Claire and Donna, who do not. Dr. Jones knows that blasting the exam room with a powerful, new inhaled chemotherapy will cure Anne, but cause at least one of Betty, Claire or Donna serious illness. Unfortunately, the exam room door is stuck; Betty, Claire and Donna cannot leave the exam room! Moreover, this is Dr. Jones's only opportunity to treat Anne. A decision must be made straightaway.

Should Dr. Jones bombard the room? I say Dr. Jones should not bombard the room, despite that it would lead to more good than harm overall. One obvious justification for this verdict is that bombarding would breach a core principle of ethical medical practice: non-maleficence (Beauchamp and Childress 2013).

It is worth spelling out, very carefully, why this is the case. We saw above that a core principle of medical ethics is non-maleficence, or to do no harm. Interpreted literally, of course, the principle is implausible. A surgeon who carefully slices open a patient "harms" her, but it would be odd to

think her actions *prima facie* unethical. Rather, a better interpretation of non-maleficence would read the principle as an injunction against imposing "medically unnecessary" harms. For example, if cutting off my finger is the only or least harmful way to save my life, then doing so is "medically necessary". But if cutting off my finger*nail* would also save my life, then cutting off my entire finger would be "medically unnecessary." Or suppose that cutting off *my* finger is the only or least harmful way to save *your* life. This would also be "medically unnecessary," because the notion should be indexed to helping the specific patient harmed by that intervention. So, here is a first pass at understanding "do no harm":

> Do No Harm (DNH): a physician is under a *prima facie* obligation not to cause medically unnecessary harm to her patients

Look again at Individual Doctor. Since bombarding the room causes "medically unnecessary" harm to Betty, Claire, and/or Donna, doing so breaches non-maleficence and is *prima facie* impermissible, even if the "gain" to Anne is greater than this "loss" to Betty, Claire, and/or Donna. Of course, the *prima facie* clause entails that DNH can be overridden by other considerations such as beneficence.[21] While it may be all-things-considered permissible to make Betty slightly queasy to cure Anne's cancer, it is much more difficult to justify making Betty seriously (but not life-threateningly) ill to cure Anne. Characterizing precisely when these considerations outweigh non-maleficence is a tricky, context-dependent matter. However, for non-maleficence to have bite, it cannot simply be overridden any time an action improves overall well-being.

---

[21] This formulation of nonmaleficence as a *prima facie* principle follows Beauchamp and Childress's (2013) characterization. After all, there are some situations in which an intervention is a net bad for an individual but the harm is "medically necessary," in the sense that the intervention would prevent harm to others. An example might be a disease carrier required to undergo painful treatment to ensure that she does not transmit the disease to others. However, it should be noted that there is an important difference between screening and cases of transmissible disease. When whether I am treated affects the health of other individuals too, then it becomes more plausible to impose interventions on me, even if they are strictly speaking medically unnecessary for me, because I pose a threat of harm to others. The spirit of this idea dates back to Mill's (1869) harm principle. But screening is very different. Whether I am (unnecessarily) treated does not affect whether others are benefitted or harmed. Hence, it makes more sense, when thinking through the permissibility of harm imposition in screening, to rely on a notion of "medically unnecessary" harm indexed to the specific patient being tested. Thompson (2017) makes a related point in a discussion of preventative public health measures, arguing that there is an important distinction between policies that involve solely intra-personal trade-offs versus those that involve both intra- and inter-personal trade-offs.

With Individual Doctor and DNH in our minds, consider the following actual health practice:

> Population Screening
> Over 10 years, 10,000 women are triennially screened for breast cancer. 400 women are diagnosed with breast cancer. For 340 of these women, screening makes no difference in outcome (285 would have survived regardless, 55 would have died regardless). Of the remaining 60 women whose outcomes are influenced by the programme, 45 women will be overdiagnosed and overtreated, and 15 women will have their lives saved.[22]

Should Population Screening be implemented? In this scenario, some women are benefitted by early detection; other women are harmed through overdiagnosis and overtreatment. But notice something peculiar. Individual Doctor, on the face of it, looks suspiciously similar to Population Screening, the latter simply being on a larger scale. In order to save some people, like Anne, we must impose harms on other people, like Betty, Claire, and Donna. Yet if Individual Doctor is impermissible because it violates non-maleficence, then it seems that Population Screening is also impermissible, because it likewise imposes medically unnecessary harms. If we take DNH seriously, then it seems to follow that Population Screening should not be implemented.

Of course, there are differences between the cases. One concerns clinical decision-making and the other policymaking; typically, participants give informed consent to screening; and in screening, it is not known in advance who will benefit and who will be harmed. Section 2.3 will elaborate on these differences, but for now the point is more modest: in many ways, the harms imposed in Individual Doctor look similar to the harms imposed by Population Screening—that is, in order to benefit some individuals, medically unnecessary harms must be imposed on other individuals. Of course, there is an important dis-analogy between the cases concerning the "identifiability" of the victims. I agree. Section 2.4 will elaborate on this difference in arguing for a novel interpretation of non-maleficence in *ex ante* terms.

[22] These estimates are roughly accurate. According to the NHS: "Overall, for every 1 woman who has her life saved from breast cancer, about 3 women are diagnosed with a cancer that would never have become life threatening" (https://www.nhs.uk/conditions/breast-cancer-screening/why-its-offered/). This is all that is needed to get my argument going.

This raises a dilemma. We seem forced to either declare all screening programmes unethical or to reject non-maleficence. This is a rather shocking conclusion! And neither horn of the dilemma seems to be a palatable bullet to bite. The conclusion that all screening programmes are ruled out by DNH is unacceptable. After all, the reasoning here threatens to generalize broadly. It is not merely screening that gets ruled out. Almost any preventative health programme with risks, when applied to a large population, will harm some individuals more than it benefits them. Declaring nearly all such health policies, from vaccinations to water fluoridation, unethical is a very high cost to pay. This is unlikely to be right bullet to bite.

The latter option is also unacceptable. In my view, this is because DNH derives from two, more fundamental moral distinctions. The first concerns the asymmetry between doing and allowing harm: all else equal, it is harder to justify actions that harm people than actions which merely fail to benefit people. Making Betty seriously ill to save Anne is harder to justify than allowing Betty to suffer serious illness, because you only have the resources to save Anne. The second concerns the asymmetry between intrapersonal justification (justifying harms with benefits to the same individual) and interpersonal justification (justifying harms with benefits to different individuals). All else equal, it is harder to justify harming Betty to cure Anne than harming Anne to cure her own cancer. Betty is not just a unit in the greater good. Treating her that way ignores the "separateness of persons" (Rawls 1999). One might doubt the ethical relevance of these asymmetries, or think that DNH can be operationalized in other ways. However, even if DNH stems from different moral underpinnings, the crucial point here is that DNH is not merely a piece of empty tradition. By constraining physicians in their duty to promote health, non-maleficence operationalizes constraints that are grounded in the rights and interests of individuals. We should not give it up easily.

At this point, it pays to clarify how these claims relate to the existing literature on non-maleficence. Gillon (1985) argues that while non-maleficence is undoubtedly an important moral principle, the injunction to "*First (or above all)* do no harm" is implausible (emphasis added). In other words, Gillon is skeptical that there is a necessary priority of non-maleficence over beneficence. To support this point, he evaluates Phillipa Foot's (1980) claim that "other things being equal, the obligation not to harm people is more stringent than the obligation to benefit people." Gillon

concurs with Foot that, at first glance, this claim appears plausible. While we have a duty not to harm all other people, we do not have a duty to benefit all other people. After all, the latter duty is impossible to fulfill. However, even if this is right, Gillon (1985) notes that it does not follow that non-maleficence has any necessary priority over beneficence: "all that follows is that the scope of non-maleficence is general, encompassing all other people, whereas the scope of beneficence is more specific, applying only to some people" (130).

My claims here are consistent with Gillon's argument. As I suggested above, it is implausible to think that the injunction to *first* do no harm should be interpreted literally. Almost every medical intervention poses risks of harm to individuals. And I agree with Gillon that non-maleficence is not "an absolute principle; it does not necessarily have priority in cases of conflict with other moral principles" (131). My point here is more modest: that non-maleficence is a principle with moral valence in the ethics of screening. In defense of this idea, note that intensive chemotherapy or preventive surgery are closer to having one's finger chopped off than a fingernail removed. Imagine that a rogue doctor sliced off peoples' body parts while under sedation because, mysteriously, this practice benefits other patients. Of course, as Gillon notes, non-maleficence needs to be considered in the context of coexisting obligations, such as beneficence. I agree. The takeaway here, though, is that this balancing of principles does not solely involve determining whether the rogue doctor's actions did more good than harm overall. After all, even if more good than harm results overall, this does not entail that non-maleficence is a hollow principle nor that beneficence outweighs non-maleficence in this particular case.

In their seminal text, Beauchamp and Childress (2013) echo these points. For their example, a surgeon could save two innocent lives by murdering a prisoner on death row to harvest their organs for transplantation. Even if this would lead to more good than harm overall in this particular circumstance, it is not morally defensible. In other words, "in some cases, nonmaleficence overrides beneficence, even if the best utilitarian outcome would be obtained by acting beneficently" (Beauchamp and Childress 2013, 115). Like Gillon, Beauchamp and Childress are skeptical that non-maleficence has any necessary priority over other bioethical principles. Again, I concur with this point. What the DNH dilemma in this section brings out, however, is that even if a screening programme leads to more good than harm overall, there are two ethical principles at

stake, and we need to be careful not to reduce our choice to thinking as if it is all about one principle solely concerned with aggregate benefits and harms.

The considerations above speak to why nonmaleficence is an important principle of medical ethics. It is worth re-emphasizing, however, that my claim is not that nonmaleficence is a "special obligation" in medicine that has absolute priority over other principles such as beneficence; following Gillon as well as Beauchamp and Childress, the claim is simply that nonmaleficence *is* a principle that is ethically important in medicine. As I discuss in more detail below, current discussions of screening appear to ignore nonmaleficence *entirely*, in that appraising screening is simply a matter of determining whether a programme reduces net morbidity or mortality. Yet, as we saw in the examples above, we cannot simply assume that if a programme is "effective" at reducing net morbidity or mortality, then it is thereby permissible.

Of course, one might deny that screening should be thought of as a medical intervention at all. Instead, one might think that screening is more analogous to other public policy interventions, like taxation. Were this the case, then the scope of non-maleficence may not apply to screening, insofar as social policies may not be guided by the same ethical principles as clinical medicine. As Chapter 1 discussed, screening programmes appear to be a hybrid of clinical medicine, focused on the testing of individual patients, and public policy interventions, with the organization of that testing focused on populations. Despite this, however, there are good reasons to think that screening programmes qualify as medical interventions. In support of this idea, consider a single patient undergoing a Pap smear for cervical cancer. It would be odd to deny that this screening test is a medical intervention. Why, then, would the fact that millions of women undergo Pap smears in a screening programme alter the nature of Pap smears as a medical intervention? A screening programme changes how Pap smears are offered to women, but there is no reason to think this transforms the medical nature of the test.

Another way to justify the claim that screening is a medical intervention comes from Edward Pellegrino's (2001) work on the "internal morality" of clinical medicine. According to Pellegrino, the clinical encounter is the central moral phenomenon of clinical medicine. On this picture, the face-to-face encounter between physician and patient is the starting point for an ethics of medicine; the nature of this clinical encounter is what gives "moral force to the duties, virtues and obligations

of physicians *qua* physicians" (559). While screening programmes in the NHS initiate the medical encounter with a letter inviting individuals to screening, the actual experience of undergoing the screening test involves a clinical encounter between health professional and patient, in which the individual is seeking assistance from the health professional within a medical setting. In other words, the experience of undergoing screening falls squarely within what Pellegrino calls the "clinical encounter." Thus, if Pellegrino is right that the clinical encounter underwrites the crucial "phenomenon" of medicine, and if screening programmes involve such a clinical encounter between physician and patient, then it seems plausible to view screening as a medical intervention. As such, if I have argued correctly that non-maleficence is an important ethical principle in screening, then it appears to be a specific kind of overlooked principle in discussions of screening concerning how to think through the ethical permissibility of a programme.

The tension between screening programmes and non-maleficence remains, then, and neither horn of the dilemma is attractive. Where to from here? The best way forward is to argue that the dilemma is illusory. In the next section, I will consider and reject three ways to dehorn the dilemma. We cannot deny that screening imposes harm, nor claim that DNH is irrelevant to policy-making decisions, nor can we appeal to consent to dehorn the dilemma. These considerations set the stage for my preferred solution to the tension between screening and DNH, which I develop and defend in Section 2.4.

2.3 Three Objections

Denying Harm

The first wriggle denies that screening causes harm. Recall from the previous chapter a curious feature of screening: while we have population statistics about overdiagnosis rates, it is nearly impossible to assess when overdiagnosis has occurred in a particular individual. Establishing this individual-level claim requires access to an epistemically closed counterfactual: what would have happened had we not intervened. Further, many (overdiagnosed) individuals are grateful that screening led to (unnecessary) treatment, precisely because the harm is "invisible," thinking instead that screening benefitted them. Raffle and Gray (2007) note that this can often lead to a 'screening popularity paradox': "The greater the harm through overdiagnosis and overtreatment

from screening, the more people there are who believe they owe their health, or even their life, to the programme" (68). Hence, one might think that because the harms are imperceptible, they are not really harms at all—a bit like applying a "what you don't know can't hurt you" attitude to screening.

However, while the way in which overdiagnosis harms people is peculiar, it does not follow that screening does not harm people. If a company knowingly sells a carcinogenic product, we can be certain that the population incidence of cancer in its users will increase, but it would be nearly impossible to point to any specific person and say *her* cancer was caused by the product. Legal puzzles aside, people are still harmed by the company's actions. Moreover, if I cut off your finger to save your life, when only taking your fingernail would have been medically adequate, then I have still harmed you, even if you are gushing with gratitude. The harms of screening may be epistemically opaque and difficult to pinpoint precisely. The harms of screening may be easy to overlook and prone to misinterpretation, but this does not mean there are no harms.

Indeed, were this objection right and the harms of screening really could be dismissed so easily, then screening should be far less controversial as a public health practice. The benefits of screening may be slim, of course, but if there are no downsides then it might seem straightforward to still offer screening. As Geoffrey Rose (2008) notes in his "fundamental axiom" of preventive medicine: "a large number of people exposed to a small risk may generate many more cases [of disease] than a small number exposed to a large risk" (59). To put the point differently, even interventions with small effects can achieve meaningful overall health benefits when targeting a large population. So, even if the benefits of screening are slim, screening would appear to be easily justifiable in the absence of harms. But alas, as Chapter 1 discussed, screening is intensely controversial. In large part this is because of the harms of screening. Denying these harms is not a viable strategy to dehorn our dilemma.

Denying the Policy-Level Relevance of DNH

A second possible wriggle contends that DNH should guide decision-making in clinical but not policy contexts. To give this idea some force, consider how various principles that seem fitting in clinical ethics lose plausibility when we move to social decision-making. It seems permissible, and

maybe even laudable, for physicians to follow the "rule of rescue" and do everything to help the patient before them. At the policy-level, however, decisions must be sensitive to opportunity costs in resource allocation and broadly consequentialist reasoning may be appropriate (Goodin 1995). By extension, while the "net benefit" cannot override DNH and justify blasting the room in Individual Doctor, this reasoning loses water in Population Screening, which may be justified by appeal to consequentialist calculations. The justificatory process in our cases are dis-analogous.

To give this idea some force, one might observe that the National Institute for Health and Care Excellence (NICE), when deciding how to allocate a limited budget, appeals to principles in line with maximizing overall population outcomes—the institution seems to do this, for example, in their calculations of cost-effectiveness, which strive to maximize overall quality-adjusted life years (QALYs) achievable with their budget. NICE's decision-making process is contestable (Kamm 2015), of course, but suppose that it is at least broadly sensible. Would it not also, then, be sensible to aim to improve overall population outcomes in screening?

This is a very tempting thought. But I am not persuaded. There is an important dis-analogy between the decisions faced by NICE and by screening policymakers. NICE is deciding whom to help, given a fixed budget. In this decision-context, it may be perfectly sensible to maximize overall population outcomes in terms of QALYs. But recall from above that there is an asymmetry between doing and allowing harm, which underwrites the non-maleficence principle. If NICE decides not to fund a cancer drug which is not cost-effective, people in need of that treatment will foreseeably be harmed. NICE allows harm to befall certain people. But this is very different from screening, which involves doing harm to some to help others. In order to save Anne, screening imposes harms on people like Betty, Claire, and Donna. In situations that involve the imposition of harm and not merely the allowing of harm, like screening, a focus on overall population outcomes obscures the moral valence of the doing/allowing harm distinction. I say this is a moral mistake.

Still, though, it is consistent to uphold the doing/allowing harm distinction and to deny the policy-level relevance of DNH. This second wriggle needs to rejected more directly. I think this can be done by reflecting on the following case:

GOD'S EYE VIEW (GEV)

A policymaker is considering implementing Population Screening. After careful reflection and deliberation, she comes to the conclusion that implementing the programme confers a net benefit. But then, through a divine intervention, God provides her with supernatural foresight. SCREENING will save these 15 women: Anne Jones, Betty Smith, and so forth, but overdiagnose these 45 women: Alice Clarke, Barbara Williams, and so forth. This identifying information cannot be used to better target the programme to only benefit the 15 and avoid harming the 45.

Should GEV be implemented? I say it would be immoral to go ahead with screening in GEV. This is because GEV looks straightforwardly like a case of harming some people to help others. The case looks eerily similar to the classic bioethical dilemma in which you, the surgeon, have five patients in serious need of different organs and one healthy patient who could provide them, if you harvested his functional organs without consent. Most people claim that it is morally impermissible to harvest the organs of one healthy patient to save the other five. GEV is impermissible for similar reasons—it is impermissible to unnecessarily harm some people in order to help others in a medical context. I argued above that this distinction between intra- and interpersonal justification underpins DNH. If this is right, and if GEV is impermissible, then it seems that DNH carries moral valence even at the policy-level.

Appealing to Consent

A third wriggle appeals to the notion of informed consent. Clearly, it would be problematic to force participation in screening. Individuals should make an informed choice about whether to get tested, and this fuels one ethically salient dis-analogy between Population Screening and GEV: in Population Screening, we might reasonably infer hypothetical consent or conceivably acquire the actual consent of everyone. In GEV, we cannot reasonably infer hypothetical consent from those overdiagnosed and seeking their actual consent seems disingenuous. This may make a moral difference because of *volenti non fit injuria*: when people consent to some course-of-action in knowledge of the risks involved, they have no complaint if those risks eventuate (Feinberg 1986). Thus, one might argue that even if a screening programme leads to unnecessary harms, it may still be permissible either if we can reasonably infer hypothetical consent or if we can ensure that each individual provides actual consent. This strategy has the advantage of maintaining our normative

reactions to the above wriggles—conceding that screening imposes harms and preserving DNH as policy-relevant—but denying that screening is thereby unethical.

Unfortunately, this approach is wrong-headed. First, I am unconvinced that hypothetical consent underwrites the ethical difference between GEV and Population Screening. Even if, for some bizarre reason, we could reliably infer the consent of all individuals in both scenarios, GEV still seems harder to justify, because it uses the overdiagnosed as a means to saving the lives of others. This straightforwardly violates the "separateness of persons" in a way that Population Screening does not, and seems to better explain our normative reactions to the cases than hypothetical consent. Second, actual consent does not fare any better. In practice, there are various challenges to securing informed consent, since it requires understanding of the risks involved to be valid. Generally, though, we know that the presentation of risk affects the resulting perception of risk (Tversky and Kahneman 1981, Chwang 2015). And in the case of screening specifically, these issues are compounded by the "dread" many feel about cancer (Usher-Smith *et al*. 2017, 2018), which may hamper judicious reasoning and comprehension (van den Bruel *et al.* 2015). While good health professionals may ensure that consent is adequately informed, I am doubtful that consent can carry the moral burden required to mitigate complaints of harm in practice, given problems of framing and ignorance.

Third, and most importantly, appealing to consent does not properly respond to the concern that Population Screening violates DNH. Fully informed consent may mean that harmed individuals have no complaint, but this does not obviously justify the imposition of that harm in the first place. By analogy, imagine Alana knows both that Adam would consent to a boxing match and that she would beat him to a pulp. Adam may properly consent to the match, get pummelled, and have no complaint against Alana for suffering the resulting injuries. But why would Adam's (actual or hypothetical) consent to the match justify Alana requesting a match? It seems odd for Alana to make this offer *and* to claim that her choices are guided by a concern to avoid doing harm. Our initial dilemma remains.

2.4 *Ex Ante* Do No Harm

Here is where we stand. Section 2.2 argued that there is a dilemma between screening programmes and DNH. Section 2.3 then considered three ways to dehorn this dilemma—by denying the harms of screening, denying the policy-relevance of DNH, and by appealing to consent. These were all rejected. To motivate what I consider the best way to wriggle out of the dilemma, it is helpful to expand on why the appeal to actual consent above is inadequate. A focus on actual consent is both too strong and too weak to resolve the tension between screening and non-maleficence. It is too strong because ensuring that consent is genuinely informed is extremely hard. And it is too weak because, even if genuine informed consent is provided, this does not straightforwardly justify making the offer at all. Nonetheless, a focus on consent does gesture at something important, namely, that we should think in terms of the *ex ante* perspective of affected individuals. Whereas GEV and Individual Doctor, if implemented, both reduce the prospects of some below their *status quo*, adopting the policy in Population Screening conceivably might improve the prospects of each relative to the *status quo.* In this section, an *ex ante* reading of DNH is developed according to which this difference in prospects is ethically salient. I argue this is the best way to dehorn our dilemma.

Consider two possible formulations of DNH:

> *Ex ante* DNH: a physician is under a *prima facie* obligation never to perform an action which worsens an individual's (medically relevant) prospects
>
> *Ex post* DNH: a physician is under a *prima facie* obligation never to perform an action which causes medically unnecessary harm

Ex ante DNH is concerned with prospects, where a prospect is the utility or associated well-being of all the things that might happen, weighted by the chances each will occur.[23] By contrast, *ex post*

[23] In more technical terms, *ex ante* DNH is concerned with von Neumann-Morgenstern's expected utility, where a prospect has higher expected utility or well-being for a person just in case it would be preferred after calm deliberation with all relevant information, whilst attending to that person's interests only. Different prospects have equivalent expected utility when such deliberation yields indifference between the two prospects. However, it is worth flagging that the *ex ante* perspective is neutral on what well-being consists of. For example, one might think that well-being consists of nothing other than hedonism, yet still think that idealized preferences reliably track

DNH is concerned with actual outcomes, with whether a physician actually imposed medical unnecessary harm on her patient. To illustrate the difference, imagine two doctors, Dr. Ex Ante and Dr. Ex Post, who both have a patient with the flu. Dr. Ex Ante acts in her patient's expected best interests by prescribing a drug which, given the evidence, is very unlikely to have serious side effects for her patient. While the drug improves her patient's prospects, the side-effects tragically occur. Dr. Ex Post prescribes her patient a drug which, given the evidence, is very likely to have serious side effects for her patient. While the drug lowers her patient's prospects, by a stroke of luck the side effects do not occur and Dr. Ex Post's patient is cured. In these scenarios, Dr. Ex Ante respects *ex ante* DNH but violates *ex post* DNH. By contrast, Dr. Ex Post respects *ex post* DNH but violates *ex ante* DNH.

Ex ante DNH seems to be the better formulation. Here are three reasons why this is the case. First, it better tracks our intuitive judgments in the cases above. It seems, in a straightforward way, that the physician violating *ex ante* DNH acts improperly, whereas the physician violating *ex post* DNH does not. Dr. Ex Ante can say to her patient: "Look, to the best of my knowledge, this drug is in your best interests." Dr. Ex Post cannot make this claim in good faith. While a defender of *ex post* DNH might object that our intuitions track judgments of blame and responsibility, rather than permissibility *per se*, this seems off the mark. Any non-maleficence principle is intended for use in practical contexts, and it is an advantage that *ex ante* DNH captures these backward-looking attributions of blame and responsibility (Chalmers 2011).

Second, *ex ante* DNH better accounts for the underlying ethical asymmetries that I argued motivate the non-maleficence principle. For example, it is better positioned to explain the asymmetry between doing and allowing harm. Contrast, for example, the following variants of Individual Doctor involving only Anne and Betty. Dr. Jones decides to blast the room and make Betty seriously ill to save Anne. But by a stroke of luck, Betty does not get ill and Anne is saved. This is very different from the following: Betty is already seriously ill and Anne has life-threatening cancer. Dr. Jones has the dosage to cure only one. She gives it to Anne. By a stroke of luck, Betty regains her health without any medication.

---

what hedonism entails is the good life (Otsuka and Voorhoeve 2009).

If we take the doing and allowing harm distinction seriously, then Dr. Jones's actions in the first case seem harder to justify. It is harder to justify harming Betty to save Anne than merely failing to help Betty to save Anne. However, *ex post* DNH does not seem well-positioned to explain this asymmetry. In both variants of Individual Doctor, no medically unnecessary harm is caused, and so *ex post* DNH is silent on any moral difference between the scenarios. But this overlooks the fact that Dr. Jones's actions in the first case seem objectionable in a way that are not in the second.

One might think that a difference in *motives* explains the contrast, but this cannot be the full story. In Individual Doctor, Dr. Jones may not intend to harm Betty, Claire, or Donna to save Anne, but this would not blunt the verdict that blasting the room is ethically impermissible. Equally, a weaker interpretation of *ex post* DNH where the obligation is indexed to what a physician can reasonably foresee, rather than actual harm, will not work either, because harm is foreseeable in Individual Doctor too. A better explanation is that Dr. Jones's actions lower Betty's prospects to help Anne in the first case whilst in the second case Betty's prospects are unaltered. As such, *ex ante* DNH better captures what is central in our normative reactions to the doing and allowing harm asymmetry.

Ex ante DNH is also better-suited to capture the distinction between intra- and inter-personal forms of justification. For example, it seems easier to justify making Anne seriously ill to cure her cancer than making Betty seriously ill to cure Anne's cancer. In the former, we can say to Anne that blasting her with radiation is outweighed by a gain to her. It is what a benevolent agent would want for Anne's own sake. But in the latter, we cannot offer a similar justification to Betty, saying that blasting her with radiation is what a benevolent agent would want for Betty's own sake. All we could say to Betty is that it is what a benevolent agent would want for Anne's sake. This interpersonal justification is harder to come by. As Section 2.2 argued, DNH is a *prima facie* injunction against "medically unnecessary" harms that operationalises this justificatory asymmetry.

For DNH to operationalise this asymmetry, however, it must not undermine acceptable forms of intrapersonal justification. In situations of certainty, as above, both *ex ante* and *ex post* DNH

concur that harming Betty to save Anne is *prima facie* objectionable. But what about situations of risk and uncertainty? What about when blasting Anne may or may not cure her cancer? This is where the *ex ante* and *ex post* readings can diverge. The problem for *ex post* DNH is that it seems ill-suited to account for the uncertainty inherent in actual clinical practice. If exposing Anne to radiation is the only and best way to improve her chances of surviving cancer, then radiotherapy is what a benevolent agent should want for Anne's own sake. Yet if the radiation turns out ineffective, making Anne seriously ill without prolonging her life, then the treatment ends up causing "medically unnecessary" harm. By the lights of *ex post* DNH, then, the treatment was *prima facie* objectionable. This is deeply counterintuitive. And this problem does not arise for *ex ante* DNH, because the notion of prospects is contingent on the occurrence of particular states of the world, weighted by their likelihood. It is therefore better attuned to the normative significance of uncertainty. While it is tragic if the radiation does not cure Anne, it does not follow that the treatment harmed her in any objectionable way. Not all "medically unnecessary" harms are morally problematic, a fact to which the *ex post* reading is insensitive.

The case for *ex ante* DNH is made even stronger by returning to our original examples. For while *ex post* DNH entails that acting in all of Individual Doctor, Population Screening, and GEV is ethically problematic because they involve the imposition of medically unnecessary harms, *ex ante* DNH (rightly) concurs on Individual Doctor and GEV but not necessarily on Population Screening. Individual Doctor and GEV, if implemented, lower the prospects of some individuals. But it is perfectly conceivable that Population Screening improves the prospects of each affected individual. Hence, *ex ante* DNH offers the most viable route to dehorning our dilemma. We can have both screening and non-maleficence. This is a great relief.

At this point, a clarification about the normative strength of the *ex ante* DNH principle is in order. One might worry that because of the *prima facie* clause, the *ex ante* DNH principle is an excessively modest position insofar as it is merely a desirable feature of screening. However, this would be too quick. To say that a principle operationalizes a *prima facie* duty is not to say that it is "merely" a desirable feature of screening programmes. *Ex ante* DNH is not offering an account of when a screening programme is right or wrong, all things considered. That would be implausible. For example, a screening programme that raises the prospects of all may still be all

things considered impermissible if it is exorbitantly costly to implement. Rather, *ex ante* DNH is identifying *prima facie* moral reasons that speak against medical interventions which lower the prospects of some affected. As such, it is possible that a screening programme is *prima facie* objectionable, because it violates *ex ante* DNH, but still permissible all things considered. Conversely, as we just saw above, it is possible that a screening programme satisfies *ex ante* DNH yet still is impermissible all things considered.

What is important here is that the *ex ante* DNH principle is identifying reasons that speak for or against a programme that are not present in contemporary discussions of screening. That is, many commentators worry about the harms of screening (e.g. Baum 2013), yet their implicit verdict that screening is problematic fails to separate these worries from concerns about "effectiveness"—in other words, commentators implicitly assume that harms matter only insofar as they are not "balanced out" by benefits. By contrast, the *ex ante* DNH principle, which I suggested derives from fundamental ethical asymmetries, grounds these harm-related worries in an ethical principle which I have argued is independently plausible, and which underscores the thought that harms matter in some way not captured just by looking at the benefits. In addition, the normative strength of the principle can be understood by analogy to the ethical obligation to acquire informed consent. When it is claimed that we should seek informed consent from patients before operating on them, there are bound to be cases which would make it implausible to think of informed consent as a "necessary condition" for permissibility—for example, cases where treating a patient without her consent is the only way to prevent a viral epidemic. The point of the "prima facie" clause, however, is to capture the sense that there is a weighty, default consideration in favour of acquiring informed consent in medical situations. In a similar vein, while *ex ante* DNH is not as strong as a strictly "necessary condition," it is still stronger that an interpretation of the principle as being "merely desirable."

Ex ante DNH, then, is not an excessively modest position. The moral issues implied by *ex ante* DNH can have practical upshots. Imagine policymakers must choose between implementing two breast cancer screening programmes. One does not satisfy *ex ante* DNH, because it lowers the prospects of the women invited who breastfed, but it leads to favorable aggregate outcomes. The other satisfies *ex ante* DNH, and while this second programme also leads to favorable aggregate outcomes, these aggregate outcomes are marginally less favorable than the first programme. In

this scenario, the consideration that programmes should maximize aggregate outcomes point us toward implementing the first programme, whereas the consideration that programmes should satisfy *ex ante* DNH point us toward implementing the other programme. The crucial point here is that, sometimes, the normative strength of *ex ante* DNH will outweigh the consideration to maximize aggregate outcomes. This may be the case, as in the example above, when satisfying *ex ante* DNH comes at only a small cost in aggregate outcomes. *Ex ante* DNH, therefore, is not an excessively modest position that is merely desirable in screening. In certain circumstances, it can recommend different courses-of-action for screening policy by identifying considerations not captured by simply looking at aggregate "effectiveness."

However, we are not yet in the clear. Imagine a water fluoridation programme which would lead to net population benefit, but which unavoidably decreases the prospects of a few because there is only one general water supply. Is *ex ante* DNH still too strong, by ruling out these net beneficial policies on the grounds of something comparatively abstract such as prospects? This goes to a central issue concerning the proper "currency" of health policy ethics: why care about agents' *prospects*, rather than concrete *outcomes*? Why forfeit tangible benefits for the sake of something flimsy like prospects? Indeed, the fickle nature of prospects can be pushed further. The trouble is that the notion of prospects seems to turn on how we classify individuals. How can something so dependent on our perspective carry so much moral weight? For instance, why assess individuals' chances of cancer in terms of age-alone, as opposed to age-plus-breast-feeding, or age-plus-breast-feeding-plus-genetics? We need an account of how to calculate these prospects in a normatively robust way. The next two sections take up these challenges.

2.5 The *Ex Ante* Pareto Principle

Why care about the effects of policies on individuals' *prospects* rather than their effects on *outcomes*? This section will address this challenge by considering an influential argument from Alex Voorhoeve and Marc Fleurbaey (2013). Their argument is that we should sometimes ignore the *ex ante* perspective and assess policies in terms of *ex post* outcomes. Their scruples are specifically with the *ex ante* Pareto principle, according to which "if an alternative has higher expected utility for every person than every other alternative, then this alternative should be

chosen". But their discussion is noteworthy and pertinent to this chapter, not only because it is likewise framed with breast cancer screening, but also because their argument, if successful against *ex ante* Pareto, seems to entail parallel complications for the related *ex ante* DNH principle defended above. Here, I outline Voorhoeve and Fleurbaey's general argument, sketch their apparent relevance to my claims, then defend *ex ante* DNH as a plausible account of the ethics of screening in particular and policymaking in general.

Consider a choice between Policy 1 and Policy 2:

> Policy 1: Gives an equiprobable chance of (A cured, B very ill) in State of the World 1 OR (A very ill, B cured) in State of the World 2
> Policy 2: Leaves both A and B slightly ill in State of the World 1 and 2

Suppose the utility of being cured is 1, of being slightly ill is .8, and of being very ill is .65. The different possible states of the world can be represented, in table form, as follows:

Action	Person	State of the World 1	State of the World 2
Policy 1	A	1	0.65
	B	0.65	1
Policy 2	A	0.8	0.8
	B	0.8	0.8

The *ex ante* Pareto principle recommends that Policy 1 should be chosen over Policy 2. This is because the *expected utility* of both A and B is higher under Policy 1 than under Policy 2. A 50% chance of being cured and a 50% chance of being very ill yields a higher expected utility than a 100% chance of being slightly ill for both A and B, since $(.5)*(1) + (.5)*(.65) > 0.8$.

Voorhoeve and Fleurbaey take issue with *ex ante* Pareto. On their view, "an egalitarian who rightly cares both about reducing outcome inequality and about increasing individuals' well-being" should adopt Policy 2 instead. Although Policy 1 improves the prospects of A and B more than Policy B,

we should also care about equality, and in the case above egalitarian considerations override the recommendations of *ex ante* Pareto—for a modest decrease in prospects, Policy 2 achieves a considerably more egalitarian state of affairs. From this highly stylised example, Voorhoeve and Fleurbaey generalize their conclusions to breast cancer screening recommendations. They suggest that in the choice between two screening policies, where the first is *ex ante* Pareto superior to the second, but the second foreseeably reduces outcome inequality more than the first, we may have reason to go against *ex ante* Pareto and choose the second policy. Egalitarian concerns sometimes override the importance of prospect improvement.

At first glance, the *ex ante* DNH principle defended above, which set out a *prima facie* constraint against programmes that lower the prospects of some individuals relative to the *status quo*, seems to be in trouble. This is because *ex ante* DNH appears to be a stronger version of *ex ante* Pareto—whereas the standard *ex ante* Pareto principle treats a policy being *ex ante* preferable for each affected individual as a sufficient reason to choose that policy, the *ex ante* DNH principle entails that going forward with a policy which lowers the prospects of some is *prima facie* impermissible. Put another way, the standard *ex ante* Pareto is silent on how to choose between policies in which no alternative is *ex ante* preferable for everyone i.e. when no matter which available option we choose, we will lower some people's prospects. But *ex ante* DNH implies that policies which are not *ex ante* preferable for everyone are *prima facie* impermissible. Hence, there is a sense in which the *ex ante* DNH principle entails a more demanding constraint. If Voorhoeve and Fleurbaey have good arguments against standard *ex ante* Pareto, then it seems the proposed *ex ante* DNH principle is in (even more) trouble.

In response, it is important to note that one could reject *ex ante* Pareto and still hold *ex ante* DNH. It is compatible, for instance, to recommend against a screening programme that improves each persons' prospects on egalitarian grounds and think that programmes which lower the prospects of some are *prima facie* objectionable. More substantively, whatever the independent appeal of Voorhoeve and Fleurbaey's arguments, I am sceptical they extend to the specific case of breast cancer screening. The problem is that the case-studies used to underwrite their argument only involve choices between policies that benefit (or, at least, that do not harm) different individuals, both in terms of prospects or outcomes, relative to the baseline of inaction. Any "costs" associated

with policy choice are opportunity costs, like deciding whether to help a seriously ill Betty or a life-threatened Anne. But this misrepresents the ethical issues raised by screening programmes, which I argued above directly harm some in terms of prospects or outcomes. Screening is more like *making* Betty seriously ill to save a life-threatened Anne. The *ex ante* DNH principle is intended to guide decision-making in this context of (risky) harm imposition, by setting a constraint on which policies are permissible. However, it is silent on which policy we should choose within the space carved out by that constraint, say, on the choice between two or more alternatives that improve the prospects of everyone. Accepting *ex ante* DNH is thus compatible with thinking there are egalitarian reasons to rejecting *ex ante* Pareto.

Strictly speaking, then, the "letter" of Voorhoeve and Fleurbaey's argument against *ex ante* Pareto does not actually pose a challenge for the *ex ante* DNH principle. They are addressed to different decision-contexts. In the choice between policies that lower the prospects of some (contexts of harm imposition), *ex ante* Pareto is silent and *ex ante* DNH implies that they are *prima facie* impermissible. In the choice between policies that increase or leave unaffected the prospects of everyone (contexts of distributing goods), *ex ante* DNH is silent and ex *ante* Pareto mandates the policy that most improves prospects for each. Nonetheless, the "spirit" of Voorhoeve and Fleurbaey's argument still seems to raise a tension, since it recommends a focus on *ex post* outcomes, rather than *ex ante* prospects, when assessing policy options. In the remainder of this section, I clarify how my arguments relate to these claims about the ethical (un)importance of the *ex ante* perspective.

There is a clear difference between prospects and outcomes. One might worry that a focus on prospects does not necessarily track what actually matters to us. If I undergo a prospect-lowering surgery, but the operation goes well, then it may seem odd to care that the procedure lowered my *ex ante* prospects. Nonetheless, it does not follow that prospects are entirely irrelevant to ethical reasoning. It is not uncommon to assess how well-off people are in terms which differ from those which ultimately matter to them (Cohen 1989; 2011; Arneson 1989). For example, although cultivating wealth is not the be-all-and-end-all of existence, one might argue that the state should still assess our well-being in terms of income.

Voorhoeve and Fleurbaey seem to concede this general point when they propose that a focus on the *ex ante* perspective is grounded in concerns about fairness and equality. Specifically, we may have fairness-based reasons to be concerned with the distribution of chances. They refer to a "distributive view," according to which "a given outcome inequality among people with equally strong claims to a benefit is less unfair when each person has a chance to end up better off than when the worse has no such chance because, in receiving this chance, each person receives an equal share of something of expected value" (124). So, for example, a policy in which both A and B have an equiprobable chance of being cured versus left very ill (Policy 1 above) is fairer than a policy in which A is certain to be cured and B is certain to be left very ill. This is because in the former case, on the "distributive view," an individual possessing an equal chance of being cured makes a contribution to fairness in the distributive process in a way that the latter policy, in failing to give A and B an equal chance of cure, does not.

Some philosophers, of course, reject the distributive view (Wasserman 1996). However, Voorhoeve and Fleurbaey point out that if the distributive view is right, then decisions between different policies will involve a balancing of concerns between equal chances, on the hand, and improving the outcomes of those who end up worst off, on the other hand. For example, in the table above, the relevant question becomes: "Is an outcome in which both people are slightly ill at least as good as the outcome in which one person is cured and the other very ill, and this very ill person had an equal chance at a full cure?" Balancing these concerns is tricky, but Voorhoeve and Fleurbaey suggest that, at least sometimes, we should care more about the distribution of outcomes than the distribution of chances. For example, an egalitarian should be more concerned that each has an equal slice of the pie than that each has an equal chance of the entire pie. This sounds right to me. I am happy to concede that the *ex post* perspective is sometimes more relevant than the *ex ante* for thinking about inter-personal demands of fairness or equality.

The crucial point to note, however, is that my general argument for the *ex ante* perspective operated via a separate justificatory pathway than the one Voorhoeve and Fleurbaey explore. Rather than argue that prospects are themselves the proper currency of distributive ethics, I have argued via the non-maleficence principle that it is ethically inappropriate to reduce some individual's prospects below the *status quo.* This was justified as the only way to hold the non-maleficence

principle, which I argued is grounded in more fundamental ethical distinctions, without ruling out large swathes of policy. Voorhoeve and Fleurbaey advance *egalitarian* reasons to care about the *ex post* perspective. But my argument for the *ex ante* perspective was motivated by a concern stemming from non-maleficence which, strictly speaking, is silent on issues of distributive concern.

One might still find the position defended here implausible, since it may rule out highly beneficial policies. Consider, for example, a water fluoridation policy that greatly improved population outcomes at the cost of slightly reducing the prospects of a few. Equally, one might worry that the arguments here seem ill-suited to explain other social policies. For instance, does *ex ante* DNH imply that taxation policies which lower the prospects of the wealthy are impermissible? I have two clarifications that serve as replies. First, *ex ante* DNH states a *prima facie* consideration which can be overridden by other principles to arrive at an overall judgment. If the fluoridation policy would beget monumental benefits, then the violation of *ex ante* DNH may be all-things-considered permissible. Likewise, if a screening programme greatly improves the prospects of all affected, save but a few, then it may be all-things-considered permissible to violate *ex ante* DNH.

Second, the argument here implicitly assumes that the *status quo* is itself a morally neutral baseline. In many cases, this assumption is unwarranted. It seems very plausible, for example, that the distribution of wealth and income is unjust. I take it that, in these situations, basic norms of justice outweigh considerations of non-maleficence. Hence, I do not think that taxation violates DNH. But am I too sanguine in assuming that the *status quo* distribution of cancer risk is itself morally neutral? This is a tricky question. As Wilson (2009) points out, the challenge here for public health ethics is to account for the fact that the distribution of health both affects and is affected by the distribution of other goods, such as wealth. For example, we know that there are social determinants of health and we know that lung cancer risk is higher in more deprived regions in the UK (Marmot 2017). So, if the distribution of wealth and income is unjust, as I conceded above, and if these social factors heavily influence cancer risk, then it seems that the *status quo* distribution of cancer risk is not a morally neutral baseline. How, then, can *ex ante* DNH rely so heavily on such a morally tenuous reference point?

This is a very serious problem indeed. But in my defence, it is worth noting that the calculation of prospects can at least try to account for these social complexities. I will say more on this in the next section. What is more, it is worth pointing out that nobody tries to justify mass population screening as a tool for rectifying background injustices. For example, it turns out that women living in the highest quintile of socioeconomic status in the United States have twice the rate of breast cancer diagnosis as women in the lower quintile, even after all the usual breast cancer risk factors were controlled for (Welch and Brawley 2018). Since high socioeconomic status is an unlikely independent risk factor for breast cancer, a better explanation seems to be that women in higher socioeconomic status are more connected to health care and more likely to undergo screening for breast cancer. Suppose that the cynics of screening are right, that breast cancer screening leads to overdiagnosis with minimal mortality reduction. Then one might think that, in the context of the United States, breast cancer screening can be viewed as a vehicle to level out income inequality, in which well-off women have their wealth reduced by having to pay more healthcare costs. But clearly this would be an inappropriate justification for breast cancer screening.

At this point, it is worth spelling out where my arguments sit in the broader philosophical landscape. There is a clear link between aggregative consequentialism and the *ex ante* Pareto principle. If Policy 1 is *ex ante* Pareto superior to Policy 2, then (given standard assumptions) Policy 1 will also have better outcomes in terms of aggregate well-being. Harsanyi (1955), in a famous theorem, uses this link to justify utilitarianism on the basis of what rational, non-risk-averse agents would choose behind the veil of ignorance. It is no surprise, then, that those who deny simple aggregation on egalitarian grounds, such as Voorhoeve and Fleurbaey, also deny *ex ante* Pareto, given the familiar critiques that utilitarianism neglects considerations of equality.

The arguments above relate to these debates in a complex way. I defended a principle akin to *ex ante* Pareto not because of a concern with aggregate outcomes, but as a counterpoint to a focus on aggregate outcomes. Viewed this way, there is an overlap and divergence between my approach and that of Voorhoeve and Fleurbaey. The key similarity is that both our approaches rule out some policies which would foreseeably improve net population well-being. We have a common enemy: claims that social policies should solely aim at maximizing aggregate well-being. The key contrast is how this opposition is justified. Whereas I grounded the objection in a person-affecting principle

of non-maleficence, Voorhoeve and Fleurbaey derive their view from impersonal concerns about distributive outcomes. This blindness to population-level outcomes, in my view, is a virtue of *ex ante* DNH.

To conclude this section, I will clarify the following concern about the exposition of my arguments. Since *ex ante* DNH is cashed out in terms of expected utility, one might wonder why the principle was not articulated, at the outset, in terms of expected utility rather than in terms of non-maleficence. The reason was to avoid confusion with a similar but very distinct principle that also relies on expected utility, namely, the *ex ante* Pareto principle discussed in this section. While both principles rely on the *ex ante* perspective, *ex ante* DNH is focused on a different decision context, one in which opportunity costs are not the only costs. After all, choosing between population-level policies is not always a matter of choosing whom to help. While there is a great deal of literature on which principles should govern the distribution of resources, on how concerns of beneficence and justice apply at the population-level, *ex ante* DNH is intended to highlight a gap in our thinking about population-level policies, namely, those that impose harms on some as they help others. In addition, as discussed above, expected utility has deep ties to aggregative consequentialism dating back to Harsanyi (1955). Yet *ex ante* DNH is intended as a principle to caution *against* the focus on aggregate outcomes. For these reasons, non-maleficence is the superior entry point to unpacking the *ex ante* DNH principle.

2.6 Calculating Prospects

The arguments of this chapter rely heavily on the notion of *ex ante* prospects. But just what are prospects? Without a clear definition, any principle relying on the *ex ante* perspective can lead to contradiction (Mahtani 2017). This is because claims about "chances" are always relative to some body of evidence and, more generally, to some reference class.[24] This is a thorny problem. And it cannot straightforwardly be resolved by appealing to what is "out there" in the world. A physician may believe that a certain pain medication is best for Betty, but unbeknownst to him, Betty's high

[24] This is true regardless of whether one prefers a Bayesian or frequentist interpretation of probability, since all probability claims are conditional, in some sense, on the reference class used. See Hájek (2007), "The Reference Class Problem is Your Problem too". It may not be true on a propensity theory of probability, but there are other, independent reasons to reject that position.

blood pressure renders that drug ineffective. In this case, there is a worldly correlate (Betty's blood pressure, amongst other factors) against which we can assess the physician's belief.

The case of "prospects" is different. It can be true that Anne's 10-year risk of breast cancer as a 45 year-old is 5%, and that as a 45-year-old-who-breast-fed it is 2.5%, yet we cannot resolve which is Anne's "actual" risk of breast cancer by appealing to facts "out there" in the world, in the way we might resolve which pain medication is best by appealing to facts about Betty's blood pressure. While the world does impose *some* constraints on reasonable risk estimates (it must be the case that the frequency of breast cancer amongst 45 year-olds is 5%), these constraints already presuppose a particular classification (age-only). Yet this presupposition is precisely what is at stake; it is what a plausible *ex ante* DNH principle must adjudicate. After all, without a satisfactory answer, *ex ante* DNH may rule the same programme permissible or impermissible depending on how individuals are classified. For example, screening may improve every 50-70-year-old woman's prospects if we assume each has an "average" risk of breast cancer, but impermissible if we calculate risk according to age-plus-breast-feeding history, since under this designation the prospects of some will be reduced relative to the *status quo.*

So, how should we designate individuals for the purposes of screening policy? In what follows I will consider and reject three possible solutions, before defending the answer I prefer. These issues of classification will arise again in Chapter 4, where I will extend the method of calculating prospects developed here to situations of uncertainty.

One possible answer is to assess policies in terms of their expected outcomes, then assume that each individual has an equal chance of "harm" or "benefit." So, for example, the present NHS breast cancer screening programme invites women based on age only. Suppose this programme leads to better aggregate outcomes. Were we to calculate individual prospects merely by assuming each individual has an equal chance of "harm" or "benefit," then this approach would simply collapse the distinction between *ex ante* DNH and consequentialist reasoning. The problem is that this seems ethically unjustifiable when we are aware of cancer-related variation within a population. Sometimes we have good reason not to presume that each individual is "average" for a population. This seems true for the case of breast cancer. It turns out that the average 10-year

absolute risk of breast cancer in women aged 50 years in the United Kingdom is 2.85%, but women at the lowest and highest percentiles of the risk distribution are at 0.53% and 9.96% 10-year risk (Pashayan *et al.* 2018). That is a somewhat wide variation and suggests that presuming each women has an equal chance of "benefit" or "harm" is ethically problematic. I will have more to say on "particularizing" population frequencies to individuals in Chapter 4.

Another possible option is to index claims about individuals' prospects to what the participants themselves believe. In light of the worries around consent raised above, however, this would be problematic. There is widespread ignorance and confusion around cancer risk. There is also widespread confusion around the benefits and harms of screening. 98% of women in France, Germany, and the Netherlands overestimated the benefit of mammography screening 10-fold, 100-fold or more, or did not know (Gigerenzer 2009). Conversely, studies suggest the majority of women underestimate harms of screening (Hoffmann *et al.* 2015). In light of these worries, what people believe their risk to be would not be a good guide to resolve our issue.

A third possible solution is to use *all* of the available evidence. After all, many think that when assigning probabilities to one-off events, rational agents should use the narrowest available reference class for which reliable statistics can be compiled. This approach has an advantage over the "average individual" strategy above, since it correctly carves out space for *ex ante* DNH to come apart from consequentialist reasoning. A policy which improves average prospects does not necessarily improve Jane's prospects. But it faces at least two disadvantages. First, the idea of "available" evidence requires specification. If one reads it as "all the evidence ever collected," or even "all the evidence ever published," then this approach would be unfeasibly demanding. For example, a recent paper (Cintolo-Gonzalez *et al.* 2015) compared the forty-nine(!) most common models for calculating breast cancer risk: even if these estimates could meaningfully be collated into a single score for each woman affected by a proposed screening policy (a dubious assumption), doing so would be extremely difficult. If one reads "available" more narrowly as "the evidence which the policymaker possesses," then this approach may not be demanding enough. Lazy policymakers, or those situated within faulty socio-epistemic networks may use too little evidence. After all, on this latter point, what evidence is available turns on what has been researched, and what has been researched can be structured by unjust institutions. An increasing worry in the public

health community is that studies frequently fail to generate evidence relevant to members of marginalised communities (Marmot 2017).

Here, then, is the strategy I prefer: for the purposes of normative assessment, we should calculate individual risk estimates relative to the evidence which it would be reasonable to expect a policymaker to collect and use. Whether some body-of-evidence is "reasonable" is a function both of epistemic concerns (e.g. whether there is a large effect associated with a risk factor) and of social/political/moral concerns (e.g. whether it would be reasonable to expect the state to collect this sort of information). On this picture, the "correct" risk estimate is that which we calculate based on this reasonable body-of-evidence. To clarify, the claim is *not* that the state should go around calculating each and every individual's risk on the basis of the "reasonable" body-of-evidence. That is time-consuming and pointless. Rather, the claim is that we should assess policies in terms of their effects on the prospects of different groups, where our choice about how to categorise the population into groups must be guided by epistemic and ethical concerns.

To illustrate, consider the current NHS breast cancer screening programme, which triennially invites all women between 50 and 70. Suppose that this age-based policy of whom to invite leads to net population benefit or, equivalently, is in the *ex ante* interests of the average women between 50 and 70. Recall from above that while the average 10-year absolute risk of breast cancer in women aged 50 years in the United Kingdom is 2.85%, women at the lowest percentile only have a 0.53% 10-year risk and woman at the highest percentile have a 9.96% 10-year risk (Pashayan *et al.* 2018). Suppose, then, that women in the bottom percentile have their prospects lowered by attending cancer screening. Is this troubling? That depends on whether it would be "reasonable" to expect the State to identify the women in the lower percentile. If it would be reasonable (e.g. the group consists of women with more than three children), then the programme violates *ex ante* DNH. It is ethically preferable to offer screening only to women with two children or fewer, say. Yet if it would be "unreasonable" (e.g. the group consists of women with an expensive-to-detect genetic variant), then the programme is permissible by the lights of *ex ante* DNH.

This approach has a few noteworthy implications. First, the "best" estimate of my risk of cancer may differ from the "reasonable" estimate the State uses to estimate my risk. For instance, the

"reasonable" estimate of Jane's risk might exclude including the expensive-to-detect genetic risk factor, yet Jane can pay to have a genetic test that might calculate a different, more tailored estimate of her risk. While one might find this odd, there are parallels in other practices. Considers QALYs. These are tools we use to measure how good or bad some outcome is for people. Yet as we noted above, how *you* personally value some condition may differ from the QALY-measure of that same state. Maybe you disvalue losing hearing more than the average person, because of your passion for music. Just as these gaps do not demonstrate that we should not use QALYs for policy purposes, so the gap between the "best" and "reasonable" estimates of risk does not pose a particular problem, either. What it does provide, however, is an important reason to secure the actual consent of real-life participants, since what raises your actual prospects can diverge from the "reasonable" citizens'.

It is worth spelling out this point in more detail, as it helps to illuminate the relationship between nonmaleficence and an individual's beliefs about her own good. *Ex ante* DNH is intended as a principle for thinking through the ethics of population health policies, like screening. The account of how to calculate prospects based on the reasonable body-of-evidence sketched above, then, helps in addressing the following: for the purposes of policymaking, how should *ex ante* prospects be calculated? In turn, this helps in determining when a screening programme is *prima facie* ethically permissible by the lights of *ex ante* DNH. But there is a related question that must be kept distinct from these: if screening is offered, how will individuals evaluate their own prospects of undergoing screening? For example, if an individual, lucidly aware of the relevant information, weighs up the risks and benefits of a screening test and places a much greater emphasis on the benefits of screening than the harms, then it is possible that her own evaluation of her prospects will diverge from those used by the State. In this case, *ex ante* DNH does not entail that the individual is wrong about her prospects, nor that the State's calculations of the individual's prospects are definitive of what counts as an *ex ante* benefit or harm for this specific individual.

Rather, it is perfectly consistent with *ex ante* DNH that the individual knows what is in her best interests better than the State. If a screening programme is deemed all things considered permissible, then the individual's (ideally informed) evaluation of what is in her interests is what determines, at that stage, whether screening raises or lowers her prospects. This is why it is so

important that individuals are not misled about the benefits and harms of screening (Gigerenzer 2014), and why it is necessary to secure the actual consent of those invited (Forbes *et al.* 2014). However, appealing directly to these individual evaluations of the good would be unhelpful for the purposes of determining how an entire programme affects the prospects of those invited, given that there is widespread heterogeneity in how individuals trade-off the benefits and harms of screening (van den Bruel *et al.* 2015). The account of how to calculate prospects developed above, based on a reasonable body-of-evidence, is intended to address this policy issue of how to determine prospects for the purposes of evaluating an entire programme.

Stepping back, I note that there is a deep issue here concerning the relationship between "harms" and "benefits" as calculated by the State and an individual's own conception of what is prudentially good for her. In Hausman's (2015) terminology, there is a distinction between the "public" value of health, which concerns the value of a health state for the purposes of health policy, and the "private" value of health, which concerns the value of an individual's health state to that individual. As Hausman notes, it is no surprise that the "public" value will often depart from the "private" value of health. After all, policy must be sensitive to moral considerations that go beyond maximizing individual health states by taking into account, for example, issues of fairness and solidarity. I agree with Hausman that measures of the "public" value of health should be thought of as political tools to achieve the aims of health policy, whereas the "private" value of health is something that will vary from person to person, depending on their own conception of what constitutes a meaningful life. The aim of this section, however, was to develop one important normative consideration for the "public" value of health, namely the concept of prospects, by sketching an account of how to calculate prospects when evaluating a screening programme.

Second, the approach here develops the debate around identified and statistical lives (Cohen *et al.* 2015). For example, Frick (2015) argues that the psychological propensity to save one "identified" individual from certain death instead of 100 "statistical" people each facing a 1-in-100 risk may be justified. This is because the "identified" individual has a stronger claim to our assistance, because her *ex ante* risk of harm is greater than each individual of the 100. Of particular interest, Frick notes that whether individuals are "identified" or "statistical" sometimes depends on how much effort we put into identifying them. Maybe the 1 "statistical" person of the 100 who will

suffer harm can be unveiled, with some digging. How should our moral thinking accommodate this? Frick claims that we should put "as much [effort] as is reasonable" into identifying possible victims. I agree. However, I also think that establishing what counts as "reasonable" is trickier than Frick lets on. The discussion above speaks to this problem.

More generally, the case of screening underscores an issue concerning how the identified and statistical lives debate is usually framed. Often, the dialectic implicitly assumes that the decision between "identified" and "statistical" lives is simply "thrown up" by the world, without any reflection on the ways in which we sometimes must *choose* whether to take certain facts into account in policymaking. In thinking about the breast cancer screening programme, we may know that our policy will harm the prospects of members at the lowest risk of cancer. To what extent should we worry about the plight of this group? That depends on whether we could reasonably be expected to identify them. However, this means that ethical and political considerations are not only relevant to thinking about how to respond to the claim of "statistical victims;" they are central to which "statistical victims" we should recognize in ethical and political thought in the first place. When philosophers implore us to decide whether to save 100 people each facing a 1-in-100 risk of harm, this disguises a complex issue concerning how we arrived at such risk estimates in the first place. Real-world phenomena do not throw us neatly packaged risk estimates to plug into our moral decision-making.

2.7 Conclusion

We are now in a better position to answer: "When is a screening programme ethically justified?" By developing and sharpening a principle of non-maleficence for screening policy, the conclusion is that any screening programme that lowers the prospects of some affected is *prima facie* impermissible. In the course of this argument, a few other philosophical advances were made. An account of prospects was developed according to which they are calculated depending on the narrowest reference class for which it would be reasonable to expect evidence to be collected. And the relationship between these arguments and recent work in the ethics of risk was explored — both around distributive justice in Section 2.5 and the identified/statistical lives debate in Section 2.6. In the following chapter, I will explore the implications of these arguments for the concept of screening effectiveness.

CHAPTER 3

Effectiveness

3.1 Introduction

The previous chapter defended a principle, *ex ante* DNH, according to which any screening programme that lowers the prospects of some affected is *prima facie* impermissible. This principle goes some way toward articulating the ethics of screening. In this chapter, I want to explore the relationship between this principle and a debate central to disagreements about screening. The debate concerns whether screening is *effective*.

Consider breast cancer screening again. On the one hand, a UK independent review of breast cancer screening concluded "that the UK breast screening programmes confer significant benefit and should continue" (Marmot *et al.* 2013). This verdict was justified on the panel's review of the evidence of benefit (roughly, one death averted for every 180 women who attend screening) and the evidence of harm (roughly, three overdiagnosed women for every breast cancer death averted). While the NHS breast cancer screening programme is currently being overhauled, the issue concerning Sir Mike Richards, who is conducting the review, is not *whether* screening is effective, but rather if it could be made more so. He comments: "There is no doubt that screening programmes save thousands of lives every year. However, as part of implementing the NHS's Long Term Plan, we want to make certain they are as effective as possible."[25]

On the other hand, many are dubious about the merits of breast cancer screening. So Prasad *et al.* (2016) provocatively title their BMJ article: 'Why cancer screening has never been shown to "save lives",' noting that no randomized trial of screening has demonstrated a reduction in overall mortality. And Peter Gøtzsche, formerly of the Cochrane collaboration, recently titled a

[25] https://www.bbc.co.uk/news/health-46212057

presentation at the University of Cambridge, "Why You Should Usually Avoid Cancer Screening," arguing fervently that mammography screening has little to no benefit but serious and prevalent harms.[26] Further, recall from Chapter 1 that when the NHS Breast Cancer Screening Programme allegedly failed to invite a cohort of women in 2018, a letter to the *Times* co-signed by 15 medical professionals intimated it was a 'lucky escape': "The breast screening programme mostly causes more unintended harm than good, has no impact on all cause mortality, and claims of lives 'saved' are counteracted by deaths resulting from interventions" (Hawkes 2018).

Such discussions in the academy and the media gloss over an issue which is both philosophically interesting and practically important: just what is it for screening to be effective? I am not alone in noticing that the concept of effectiveness warrants philosophical attention. Philosophers of medicine, in recent years, have advanced differing views on this topic. In the context of clinical treatments, Ashcroft (2002) has argued that effectiveness is always for some end, that it references a causal capacity, and that these causal capacities supervene on physical properties. Howick (2011) argues that a "clinically effective" treatment has patient-relevant benefits that outweigh any harms, is applicable to that patient, and is the best option available (24). And Stegenga (2018) argues that an effective medical intervention must target either the causal basis of a disease or the harms of it, and ideally both.

While these arguments are illuminating, the focus of these accounts is on the effectiveness of therapeutic interventions for individual patients with disease. But screening is different, in part because it targets asymptomatic people, and in part because it targets populations. In some respects, then, screening is more akin to a water fluoridation policy affecting an entire population than to the provision of chemotherapy for an individual patient. In what follows, I will argue that this entails that the concept of screening effectiveness has an underappreciated normative dimension. The aim of this chapter is to show that concepts of screening effectiveness rely on substantive and sometimes controversial moral assumptions that are disguised behind technical facts about "mortality reduction."

[26] http://www.crassh.cam.ac.uk/events/27611

The discussion unfolds as follows. Section 3.2 introduces the central claim of this chapter, namely, that effectiveness is what Alexandrova (2018) calls a "mixed claim." Following that, Section 3.3 develops two alternative ways of understanding "screening effectiveness." One is in terms of aggregate population outcomes; the other is in terms of *ex ante* individual prospects. Section 3.4 unearths the moral presuppositions underlying these different ways of thinking about screening effectiveness. Effectiveness, I will suggest, is far more value-laden than commentators let on. Section 3.5 concludes the chapter by suggesting that the best route forward is a pluralistic view of effectiveness.

3.2 Mixed Claims

The central claim of this chapter is that screening effectiveness exemplifies what Alexandrova (2018) calls a "mixed claim." This section will explain the notion of a "mixed claim"; the subsequent sections will spell out, in greater detail, the specific ways in which effectiveness is a "mixed claim." We can begin by considering the following statement:

> *Commutes*: "Long commutes are associated with lower well-being"

What sort of claim is this? *Commutes* looks to be like many other scientific claims. There are two variables (time of commute, well-being) and the relationship between the variables is uncovered with the aid of empirical inquiry (increasing commute time → lower well-being). The structure of *Commutes* appears to be no different from the following claim:

> *Oncogenes*: "Mutating the *Ras* oncogene leads to increased cancer cell proliferation"

Again there are two variables (gene mutation, cell proliferation) and the relationship between such variables is uncovered with empirical inquiry (mutating *Ras* oncogene → higher cancer cell proliferation). Both *Commutes* and *Oncogenes* are empirical claims that are well-confirmed in

social science and cancer biology, respectively (Diener *et al.* 2008; Hanahan and Weinberg 2011). They advance a causal relationship between two variables of scientific interest.[27]

But there is an important difference between *Commutes* and *Oncogenes.* It is relatively easy to define and measure whether, in an experiment, the *Ras* oncogene is mutated. Briefly, this involves checking that the genetically modified cancer cells have an increased expression of the *Ras* protein. And it is relatively easy to define and measure cell proliferation. Briefly, this involves counting the number of cells in a pre-defined area of space, and comparing the difference in the number of cells between the control (no *Ras* mutation) and intervention group of cells (induced *Ras* mutation). But how to define and measure the well-being variable in *Commutes* is open to debate. This is because well-being has a normative component which, in turn, requires a value judgment about how it is to be defined and measured.

So, for example, in the psychological sciences, there are typically three routes to go about making this value judgment. One strategy understands well-being as a favourable balance of positive over negative emotions (Kahneman *et al*. 2004). Accordingly, this strategy equates happiness with well-being and is measured with experience sampling methods. Another strategy equates well-being with life satisfaction — that is, with an individual judgment about how one's life is going overall (Diener *et al*. 2008). In turn, this approach is measured with self-reports of life satisfaction. A third strategy connects well-being to notions of flourishing or good functioning which, in turn, refer to various aspects of the putative good life such as competence or a sense of achievement and meaning (Ryan and Deci 2001). Self-reports are used to measure these aspects of life.

Which conception of well-being should be used when examining the impact of long commutes? Answering this question cannot straightforwardly be resolved by appeal to empirical facts. It is not like simply counting the number of cancer cells in a cell culture dish, after manipulating the *Ras* oncogene into overexpression. To get a grip on the variables in *Oncogenes*, one does not need to take a stance on any normative issues. To get a grip on the variable of well-being in *Commutes*, one does. As Alexandrova (2018) notes, in the latter case, we are not simply interested in *how* long

[27] I am assuming here that the statistical association between long commutes and lower well-being is causal. This assumption will not matter for the subsequent arguments.

commutes impact happiness, or life satisfaction, or flourishing. We are also interested in understanding whether long commutes are *good* or *bad* for us, and answering this question requires a stance on which conception of well-being is most plausible. In this way, *Commutes* encodes a dimension that goes beyond merely empirical issues and bleeds into the normative.

To capture the difference between claims such as *Commutes* and *Oncogenes*, Alexandrova (2018) develops the notion of a mixed claim:

> *Mixed Claim*: A hypothesis is mixed if and only if:
> 1. It is an empirical hypothesis about a putative causal or statistical relation.
> 2. At least one of the variables in this hypothesis is defined in a way that presupposes a moral, prudential, political or aesthetic value judgment about the nature of this variable.

The important dimension of mixed claims concerns the second part of the definition. The basic idea is that a hypothesis such as *Commutes* turns on a normative judgment in one of two ways. On the one hand, a researcher might adopt a particular measure of well-being because, in her view, it better tracks what well-being actually is. A scientist ardently committed to the view that the good life consists of increasing happiness, like Kahneman *et al*. (2004), may therefore use experience sampling methods to measure well-being. She might, for instance, collect data by asking participants to report on their thoughts and feelings over time. In this scenario, there is an overt normative judgment pertaining to the correct view of well-being.

On the other hand, a researcher might follow the methods set by her discipline when measuring well-being. A scientist might, for example, gather life satisfaction reports because this is the methodological norm in her particular subfield of research. But as Alexandrova (2018) points out, in these cases the adoption of life satisfaction reports "betrays an implicit normative commitment to the validity of this research program" (424). A scientist may explain her methodological choices by appeal to the norms of her subfield, yet such norms are themselves premised on a prudential value judgment—the subfield, on pain of scientific rigour, must be premised on the thought that its methodologies are tracking well-being in some way. Accordingly, scientific research on well-

being involves mixed claims, because they depend on normative judgments—either explicitly or implicitly.

In a recent book, Alexandrova (2017) argues convincingly that the science of well-being is brimming with mixed claims. The following sections will argue that claims about screening effectiveness are likewise mixed claims. To spell this claim out in detail, the strategy will be to show that there are two very different ways of thinking about screening effectiveness. I will not defend a preference for either way of thinking here, but I will show that the choice between them implicitly relies on different moral considerations. Unearthing these moral presuppositions goes some way toward clarifying how vociferous disagreement over the "effectiveness" of screening programmes arises.

3.3 Two Views of Effectiveness

3.3.1 Aggregate Population Effectiveness

It will be useful to frame the subsequent discussion with a case-study. Consider Geoffrey Rose's "fundamental axiom" of preventive medicine: "a large number of people exposed to a small risk may generate many more cases [of disease] than a small number exposed to a large risk" (Rose 2008, 59). This axiom implies that sometimes policymakers will have a choice to make between two different screening policies:

> Prevention Paradox
> You, the policymaker, are tasked with choosing between two different screening policies. The policies are cost-neutral but only one or the other can be implemented. Here are your options:
>
> 'The High Risk Strategy': formulate policy to target those in the population at highest risk i.e. those with BRCA1 and BRCA2 mutations for breast cancer
>
> 'The Population Strategy': formulate policy to target a more substantive mass of the population at 'average risk' i.e. all women between the ages of 50 and 70

Which policy should be chosen? This is a tricky decision (John 2014). I will discuss the appeal of both in this chapter. But, for now, consider one attractive aspect of 'The Population Strategy.' A curious feature of this approach is that it generally improves overall population health more than 'The High Risk Strategy.' Even though each woman targeted by 'The Population Strategy' has a comparatively lower risk of developing breast cancer, this policy will detect many more cases of breast cancer, simply because there are so many more women between the ages of 50 and 70 than with BRCA1 and BRCA2 mutations. So, assuming screening reduces breast cancer mortality, 'The Population Strategy' will lead to better aggregate outcomes, since Rose's "fundamental axiom" entails that screening many more individuals at a small risk of disease will lead to the early detection of more cancers than focusing only on those at highest risk.

Suppose you think that 'The Population Strategy,' in the choice above, is preferable. One justification for this intuition is that 'The Population Strategy' seems to follow naturally from an independently attractive moral principle:

> *Save the Greatest Number*: all else being equal, we should choose the policy that saves the most lives as possible

So, for example, if you face the decision between saving one life and saving 10 different lives, *Save the Greatest Number* implies you should save the 10, all else being equal. Here is a more vivid case to flesh out the principle: while cruising in the Pacific Ocean, your radio informs you that there is one person stranded on an island five hours to your east, and ten people stranded on an island five hours to your west. Unfortunately, though, you have a limited and finite amount of fuel remaining. You can turn your boat east, and save one individual; or you can turn your boat west, and save ten individuals. But you cannot make both trips. What is more, you have no special relationship to any of these stranded individuals. There is no compelling reason for thinking that one of these individuals is worth rescuing more than the others. Which way should you turn your boat— east or west?

John Taurek (1977) made this sort of case famous in normative ethics. He argued, provocatively, that sometimes you should reject *Save the Greatest Number*. In this situation it is not true that you should turn your boat west and save the ten individuals. Instead, what you should do is flip a coin.

You should then decide which way to turn depending on the outcome of the coin toss.

Taurek's rationale for denying *Save the Greatest Number* rested on two ideas. The first idea was that *Save the Greatest Number*, in meshing with a consequentialist approach to ethics, implicitly denies the "separateness of persons." Recall from Chapter 2 that many philosophers argue that there is a moral asymmetry between cases in which actions affect only one person, harming her in some ways and benefitting her in other ways, and actions that affect different individuals, harming some people but benefitting other, different people. The latter case of *interpersonal* justification is harder to justify. I will say more about this below. The second idea concerned fairness. When deciding which way to turn your boat, what you ought to do is be fair to all of the affected parties. It would be unfair to the one individual, stranded to your east, to turn your boat west instead on the grounds that she is only one and the other island harbours ten individuals. To be fair to all of the stranded individuals, what you ought to do is give each person an equal chance of survival. This might involve flipping a coin, or spinning a bottle, or tossing a dice weighted according to the number of people on each island (Broome 1998).

While many philosophers are sympathetic to Taurek's rationale, few ultimately accept his conclusion that *Save the Greatest Number* should be rejected. Most philosophers think that, in the situation above, one ought to turn the boat west and save the ten individuals. If you are also in this philosophical majority, thinking that *Save the Greatest Number* is compelling, then here is a natural way to state a view of effectiveness along these lines:

> Aggregate Population Effectiveness: all else equal, an effective screening programme is one that leads to the best (health-related) population outcomes

So, for example, recall the following case from Chapter 2:

> <u>Population Screening</u>
> Over 10 years, 10,000 women are triennially screened for breast cancer. 400 women are diagnosed with breast cancer. For 340 of these women, screening makes no difference in outcome (285 would have survived regardless, 55 would have died regardless). Of the remaining 60 women whose outcomes are influenced by the

> programme, 45 women will be overdiagnosed and overtreated, and 15 women will have their lives saved.

If the choice is between whether Population Screening or no screening should be implemented, then the programme above is Aggregate Population Effective insofar as it brings about a better state of affairs than no screening. More specifically, Population Screening is Aggregate Population Effective insofar as the burdens of testing thousands of women and unnecessarily treating 45 of them is worth the benefits of saving 15 lives. This seems straightforward enough. What is more, this way of thinking meshes well with standard approaches to thinking about screening effectiveness. For example, four current or previous members of the U.S. Preventive Services Task Force recently argued that screening should be evaluated based on whether the population "benefits" outweigh the population "harms" (Harris *et al.* 2011).

Before proceeding, I will note a complexity here. There is considerable conceptual space between the "Save the Greatest Number" Principle and "Aggregate Population Effectiveness." Sometimes, the best population-level outcomes might not involve saving the greatest number. Contrast, for example, a screening programme which saves more people but has worse aggregate outcomes, with another programme that saves fewer people but has better aggregate outcomes. One way this might occur is the following: a programme may save more people but have worse aggregate outcomes if it involves unnecessarily treating (but not fatally harming) many people. Another way this might occur is the following: a programme may save less people but have better aggregate outcomes if it has a negligible impact on mortality but turns out to improve aggregate quality of life by, say, reducing anxiety about cancer.

Thus, one might be committed to Aggregate Population Effectiveness yet deny that Save the Greatest Number is the apt or only principle to underwrite it. This is an important qualification to point out. However, given the predominant focus in screening debates over whether programmes reduce overall mortality, I suggest that the Save the Greatest Number Principle is an intuitive way of unpacking the notion of Aggregate Population Effectiveness, even though it may not be the only way.

It is worth pointing out two features of this way of thinking about screening effectiveness. One is that it connects "screening effectiveness" to a utilitarian perspective, which is concerned with opting for the policy that leads to the best overall (expected) outcomes. In this way, the *Save the Greatest Number* moral principle is one way to operationalize Aggregate Population Effectiveness. Given two states of affairs that differ only in terms of the number of lives saved, for example, a plausible utilitarian axiology will rank the state of affairs where more lives are saved as superior.

The other feature is that the view implies "screening effectiveness" is a population-level property. In Chapter 1, we saw that screening programmes are sometimes understood as population interventions, by acting as a "sieve" to filter out certain individuals at increased risk. This may provide some reason to think that "effectiveness," too, should be understood as a property of populations.

If "screening effectiveness" is a population-level property, then there is an interesting advantage of this way of thinking. Notice that in Population Screening, it is only individuals that bear uncertainty, so to speak, because for each woman we do not know whether she will be overdiagnosed, or saved, or unaffected by the programme. But there is no uncertainty at the population-level. By stipulation, we know the distribution of outcomes if Population Screening is implemented. And this is often true in practice, too, because given a certain amount of statistical knowledge, we can typically predict with high certainty the distribution of outcomes if a screening programme is implemented.[28] How can we be so certain? Well, given a very large population in which individual outcomes are statistically independent, the Law of Large Numbers tells us that we can predict with very high confidence that the overall pattern of outcomes will be very similar. These assumptions hold in breast cancer screening, since a programme usually affects millions of women, and since we have good reason to think that (say) Anne developing breast cancer does not alter the chance that (say) Betty will develop it.

So Aggregate Population Effectiveness, in the context of screening, has the advantage of avoiding a sense of fickleness. When dealing with the effects of policies at the population-level, we often

[28] Of course, in practice the population statistics may, themselves, be uncertain. I set these issues aside for now. I will explore these issues, in great detail, in the following chapter.

know what we will get at the population-level—given that certain assumptions related to the Law of Large Numbers hold. Contrast this scenario with a situation in which there is uncertainty at the population-level. Imagine that there is evidence suggesting a nearby volcano might forcefully erupt, at some point, in the next week. In this case there is uncertainty at the population-level, because the fate of the population is "tied" together, as it were. If a policy recommends the city to evacuate, then everyone will presumably evacuate; if a policy recommends staying, then everyone will not evacuate. But given a certain amount of statistical knowledge—say the best available evidence suggests that there is exactly a 37.56% chance the volcano will erupt in the next week—we cannot predict with high certainty the outcome for the population, regardless of whether the city population is massive or minute. The Law of Large Numbers is no help in this situation.

Of course, even if we can predict with near certainty the overall effect of Population Screening, this is a separate matter from adjudicating whether the programme brings about a world with more good than harm. This is a hard moral problem, a matter of weighing up goods, like people's lives, that are not straightforwardly commensurable with other not so good things, like unnecessary cancer treatment. But at least we know what benefits and harms, in roughly what quantities, need to be balanced.

3.3.2 Ex Ante Individual Effectiveness

What to make of the view of screening effectiveness above? One thing to note is that, in connecting "effectiveness" to a utilitarian outlook, Aggregate Population Effectiveness relies upon a particular way of thinking about *interpersonal aggregation*. As we saw in Chapter 2, it is typically harder to justify harming Betty to prevent Anne's illness than harming Anne to prevent future illness in Anne. But Population Screening involves this former, more difficult to justify type of aggregation. The individuals who benefit from screening are *not* the same individuals who are harmed. So, you can quite reasonably take issue with Aggregate Population Effectiveness on the grounds that it seems to implicitly involve a morally contentious way of weighing up benefits and harms to different people. Judith Jarvis Thomson (1993), for example, argues that the mistake committed by utilitarians is thinking that states of affairs can be better or worse *simpliciter*. But this is not quite right. In her view, "all goodness is goodness-in-a-way" (Thomson 1993, 149). So, one

screening policy can bring about a state of affairs which is very good for Anne, but moderately bad for Betty. Another screening policy can bring about a state of affairs which is moderately good for Betty, but very bad for Anne. Yet, the most we can say when comparing these different states of affairs is the particular *ways* in which one is better than the other. Neither state of affairs is *just* better than the other.

A separate but related thing to note is that, thinking "screening effectiveness" is a population-level property seems to imply treating the "population" as some sort of "super-individual." But some philosophers deny that anything with morally significant interests consists of different people as its parts. Here, for example, is Robert Nozick on the matter:

> Individually, we each sometimes choose to undergo some pain or sacrifice for a greater benefit or to avoid a greater harm: we go to the dentist to avoid worse suffering later; we do some unpleasant work for its results; some persons diet to improve their health or looks; some save money to support themselves when they are older. In each case, some cost is borne for the sake of the greater overall good. Why not, *similarly*, hold that some persons have to bear some costs that benefit other persons more, for the sake of the overall social good? But there is no *social entity* with a good that undergoes some sacrifice for its own good. There are only individual people, different individual people, with their own individual lives. Using one of these people for the benefit of others, uses him and benefits the others. Nothing more. What happens is that something is done to him for the sake of the others. Talk of an overall social good covers this up. (Intentionally?) To use a person in this way does not sufficiently respect and take account of the fact that he is a separate person, that his is the only life he has (Nozick 1974, 32-3).

So, in adjudicating between Population Screening and no screening, Aggregate Population Effectiveness seems to require a stance on which policy is better, overall, for some abstract social entity. For example, if we think that the 15 lives saved by Population Screening outweigh the costs of overtreating 45 women and unnecessarily testing thousands of others, then we might say that Population Screening is "better overall" for the social entity consisting of the population of women affected. If the interests of this social entity were morally significant, then this would generate a reason to favour Population Screening. Nozick's point was that, by attributing moral significance

to a policy being "better overall," we are implicitly supposing that such a social entity exists. But no such social entity exists.

In light of these complications, it pays to return to Prevention Paradox and consider the other strategy. One appealing aspect of 'The High Risk Strategy' is that it seems to follow naturally from another independently attractive moral principle:

> *Concentration of Risk*: all else being equal, we should help those facing a greater risk of harm than those at a lower risk of equivalent harm, because the concentration of risk has moral valence

So, for example, if Anne faces a 90% chance of developing life-threatening cancer and Betty faces only a 30% chance of developing life-threatening, and if we can only help one or the other, then we ought to help Anne, morally speaking. *Concentration of Risk* has some philosophical supporters. Norman Daniels (2015), for instance, invites us to consider the following case in which *only* the concentration of risk varies:

> Treatment vs Vaccination
> You, the physician, have only five tablets of medicine. These can be used either to successfully treat disease (*Treatment*) in a single individual with all five tablets, or to successfully prevent disease (*Vaccination*), whereby the tablets are given, one each, to five different people. Without *Treatment*, the individual was die. Without *Vaccination*, one of the five individuals will die. So, for example:
>
> *Treatment*: Anne has the disease. With the whole dose of five tablets, she will survive. Without all five tablets, she will not.
>
> *Vaccination*: Betty, Carol, Debbie, Ella, and Francesca have been exposed to Anne. Give them each one tablet, and this preventative vaccine will save all five. Without the vaccine, one of these five individuals will not survive.

Let us stipulate that, whichever course-of-action you pursue, you will save only one expected life. This controls for issues of aggregation and any statistical complications, for instance, that by a stroke of luck more or fewer lives may actually be saved. And notice that, whichever course-of-action you pursue, you know the identities of the people you will be affecting. This controls for

some of the issues around the identical vs statistical lives bias (Cohen *et al.* 2015).

A clarificatory aside on this last point: social scientists have documented an "identifiable victim effect" (Jenni and Loewenstein 1997). Roughly, the idea is that people are willing to spend more resources to save identified lives than to save equal numbers of unidentified, "merely statistical" lives. The effect has been empirically demonstrated across many contexts (Slovic *et al.* 2004). And it has explanatory power. It is invoked to explain why we are relatively strongly motivated to rescue drowning children from nearby ponds, but far less motivated to donate money to charities that help distant strangers. It is also invoked to explain why we are relatively strongly motivated to support programmes that strive to cure illness—say, by distributing antibiotics—yet far less motivated to support programmes that prevent illness—say, by distributing vaccines.

It is important to control for this identified vs statistical lives distinction because there is disagreement over what, ethically speaking, underwrites it. For one thing, notice how opaque the identified versus statistical lives distinction can be. It may be that we do not know the identity of the relevant individuals at present, but we are in a position, with a little digging, to find out. Maybe we know of the individuals, but their identities are underspecified—say, the people attending the Los Angeles Lakers basketball game next week. Or it may be that we know a great deal about the effects of our actions, and not the particular people affected: perhaps we know that action A will greatly benefit one person in a specific way, but are unclear about how it might negatively affect another person. Or maybe we know that action A will benefit one person, but action B will yield an equivalent *expected benefit*—say, if it involved a one-in-five chance of benefitting five people, and a four-in-five chance of benefitting nobody. Or maybe we only have a very minimal form of identification, say, if all we know is that the beneficiary has a moustache.

For another thing, we might ask, in light of this documented psychological predisposition, whether it tracks anything of *moral* salience. After all, one might claim that there is no necessary connection between psychological facts and moral facts. When we see advertisements by charities that showcase a desolate puppy in need of help, we may feel the emotional tug to assist it. But it does not necessarily follow that donating to an animal charity is what morality recommends. Perhaps morality requires us to take the point of view of the universe, and to think that suffering is intrinsically bad. Perhaps, then, there are more cost-effective places to direct our donations—say,

if the amount of net suffering we can reduce by purchasing mosquito nets, on any metric of cost-effectiveness, will be far greater than donating to the animal charity, even though the puppy is adorable and precious and you want with all your heart to help it. Regardless of whether this argument advanced by "effective altruism" is sound, it raises the worry that there is space for divergence between your emotional attitudes to cases and the attitudes you should have, from a moral perspective, to such cases.[29]

In their seminal study on identified vs statistical lives, Jenni and Loewenstein (1997) examined four potential factors that could underwrite the identified versus statistical lives distinction: (1) vividness, (2) certainty of threat to identified lives and the probabilistic threat to statistical lives, (3) proportion of reference group that can be saved, and (4) ex ante versus ex post evaluation. They found that the third factor—proportion of reference group that can be saved or, in other words, the concentration of risk—was the major cause of the identifiable victim effect. So, the question to ask here, and the one raised by Daniels' thought experiment, is this: does the concentration of risk in Treatment vs Vaccination make a *moral* difference? Do you have a greater obligation to choose *Treatment* or *Vaccination*?

Daniels (2015) says this: "I believe we have a stronger obligation to treat [Anne] than to vaccinate the five others" (119). He is not alone. Frick (2013), working on a contractualist framework of morality (Scanlon 1998), writes that "I believe that individuals have a stronger claim to be protected from a harm that they would otherwise suffer with certainty than from a mere risk of suffering a harm of equivalent size" (188). So, in the case above, Frick is also in favour of opting for *Treatment*, because Anne's claim to assistance is comparatively stronger than the individual claims of Betty, Carol, Debby, Ella, and Francesca, by virtue of Anne being at greater risk of suffering harm.

Suppose, like Daniels and Frick, that you find *Concentration of Risk* to be a plausible moral principle. And suppose that, as a result of this, you think that 'The High Risk Strategy' in

[29] See, for example, Peter Singer's (1972) "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1(3): 229-243 for the canonical argument underpinning the "effective altruism" movement.

Prevention Paradox is the preferable policy, because this policy is more "effective" than 'The Population Strategy.' Here is a natural way to state your view of effectiveness more precisely:

> *Ex Ante* Individual Effectiveness: all else equal, an effective screening programme is one that improves the individual prospects of those at greatest risk of serious harm as opposed to those at lower risk of equivalent harm

This way of thinking about screening effectiveness connects "screening effectiveness" to the *ex ante* perspective, which is concerned with prospects. Recall from the previous chapter that prospects should be calculated according to the body-of-evidence it would be reasonable to expect a policymaker to collect and use. On this approach, whether some body-of-evidence is "reasonable" is a function both of epistemic concerns (e.g. whether there is a large effect associated with a risk factor) and of social/political/moral concerns (e.g. whether it would be reasonable to expect the state to collect this sort of information).

It pays to spell out very carefully how, exactly, Aggregate Population Effectiveness and *Ex Ante* Individual Effectiveness are different ways of thinking about effectiveness. John (2014) notes that cases like Prevention Paradox come in two forms. In what John calls the "absolute prevention paradox," each individual affected by 'The Population Strategy' prefers *not* to be affected by the policy. So, for example, it may be the case that each "low risk" individual reasonably prefers not to be offered screened, because the harms of false-positives and overdiagnosis are not worth the small reduction in cancer mortality. In this situation, 'The Population Strategy' does not appear to be in the *ex ante* interests of the individuals affected and yet, from the perspective of aggregate population outcomes, the policy still seems worth pursuing. After all, a small reduction in cancer mortality, when dealing with millions of individuals, can save a significant number of lives! In this situation, 'The Population Strategy' is Aggregate Population Effective but not *Ex Ante* Individual Effective—indeed, by the lights of *ex ante* DNH it is *prima facie* ethically impermissible! I will examine this relationship between non-maleficence and effectiveness more closely in the next section.

However, some philosophers argue that Rose's 'prevention paradox' is not really a paradox at all, because the costs and benefits of a 'Population Strategy' arise in different futures of the same

individual. Thompson (2017), for instance, discusses the introduction of a "fat tax" to reduce deaths from coronary heart disease, an example of a population strategy. If there is evidence from epidemiological studies that suggest a small risk reduction in heart disease associated with the consumption of fatty foods (say it prevents 1 in 50 heart attacks), then the total number of lives saved by this policy, which affects millions of individuals, will be a substantial benefit. Equally, however, this policy entails that 49 out of 50 individuals at risk of a heart attack change their diets and witness no benefit at all, at least setting aside coincidental benefits such as an improved physique. Like screening, this fat tax raises a variant of the 'prevention paradox.' The fat tax is "…a preventative measure that brings large benefits to the community offer[ing] little to each participating individual" (Rose 2008, 48).

Yet, Thompson (2017) argues that the paradoxical appearance of prevention can be dissolved insofar as the "absolute prevention paradox" merely involves *intrapersonal* trade-offs. He explains:

> There is no (direct) causal link between one person changing their diet and another person's life being saved. Changing your diet reduces the risk of death *for you*; changing your diet does not reduce the risk of death for anyone else. Any variation in the probability of your death from heart disease, that arises from changing or not changing your diet, is independent of the probability of death from heart disease for all other individuals. As a consequence, the decision to comply with the population strategy is not really an altruistic act, since compliance does not benefit other individuals (Thompson 2017, 4).

Notice, however, that even if the "absolute prevention paradox" only involves intrapersonal trade-offs from the perspective of *each individual*, it may still involve interpersonal trade-offs from the perspective of *policymakers* deciding whether a given programme is Aggregate Population Effective. And this distinction between *intra-* and *interpersonal* justification, which I argued in Chapter 2 underwrites the principle of non-maleficence, is all that is needed to get the "prevention paradox" off the ground. For example, many non-consequentialist philosophers reject that many small harms to different people outweigh one large and meaningful benefit to one person. These philosophers might claim that, morally speaking, it is better to save one life than to avoid false-positive anxiety for a hundred different people. But this is compatible with an anxiety-averse

individual, faced with a risk of false-positive result and a small risk reduction in breast cancer mortality if screened, weighing up this trade-off very differently because only her own interests are at stake. Given a disparity between intra- and interpersonal aggregation, then, the same screening policy can be Aggregate Population Effective but not *Ex Ante* Individual Effective in cases like the 'absolute prevention paradox.'

3.4 Effectiveness as Value-Laden

I have sketched two ways of thinking about screening effectiveness. Which of them is right? I am sceptical that any general argument can be given to the conclusion that one is more fundamental than the other, that one is *the* correct reading of effectiveness. This is a difficult, unresolved and maybe unresolvable problem. I feel the tug of both. In this section I want to explain my ambivalence, by way of sketching in some detail how the two views of effectiveness turn on different moral perspectives. Briefly, the point is that if you find one view very appealing (say it is Aggregate Population Effectiveness), then you are implicitly privileging a certain way of thinking about morality. If you find another view very appealing (say it is *Ex Ante* Individual Effectiveness), then you are implicitly privileging a different way of thinking about morality. This is not problematic *per se*, of course. If, as I will suggest, claims about effectiveness are "mixed claims," then it follows that *any* notion of effectiveness will presuppose a certain way of thinking about morality. But it is nonetheless important to be mindful of what we are presuming when we advance a claim with considerable health import such as "screening is effective." This section will articulate the contours of these underlying moral commitments.

I will start by clarifying the relationship between Aggregate Population Effectiveness and the *ex ante* DNH principle defended in Chapter 2. At least sometimes, *ex ante* DNH rules out screening programmes which are Aggregate Population Effective. We saw this above in what John (2014) calls the 'absolute prevention paradox.' Here is another example. There are many guidelines for screening programmes influenced by the original criterion developed by Wilson and Jungner discussed in Chapter 1. Consider the ninth principle listed by the UK's National Screening Council (NSC): "the risks, both physical and psychological, should be less than the benefits." While this principle appears closely related to *ex ante* DNH, participants in the screening debate often assume

that whether a programme meets the NSC's ninth principle hinges solely on whether it provides a net benefit for the population i.e. whether it is Aggregate Population Effective.

But the arguments from the present and previous chapters show why this is problematic. Recall the following example from Chapter 2:

> GOD'S EYE VIEW (GEV)
> A policymaker is considering implementing Population Screening. After careful reflection and deliberation, she comes to the conclusion that implementing the programme confers a net benefit. But then, through a divine intervention, God provides her with supernatural foresight. SCREENING will save these 15 women: Anne Jones, Betty Smith, and so forth, but overdiagnose these 45 women: Alice Clarke, Barbara Williams, and so forth. This identifying information cannot be used to better target the programme to only benefit the 15 and avoid harming the 45.

I argued, in the previous chapter, that implementing GEV would be morally impermissible because it violates non-maleficence. So what? You might wonder. Our focus in this chapter is on effectiveness, and it is perfectly consistent to think that an intervention is both medically effective and impermissible. Just consider a Jehovah's witness refusing a blood transfusion which is in his medical interests. Even if the blood transfusion is in his medical interests, it may still be impermissible to impose the treatment on the patient, on pain of respect for autonomy. In Pellegrino's (2001) terminology, this is because what is "medically good" for a patient may not coincide with the "patient's perception of the good."

What is noteworthy here, however, is that if you think Population Screening is Aggregate Population Effective, then you must also think the same about GEV, because both policies lead to the same overall population outcomes. The point here is not that Aggregate Population Effectiveness must therefore be rejected across the board, nor that *interpersonal aggregation* is always a moral mistake. The point is simply that a programme being Aggregate Population Effective does not entail that the programme is permissible by the lights of *ex ante* DNH. The crucial takeaway, then, is this: we cannot equate the "effectiveness" of a screening programme with the "ethically permissibility" of that same programme.

To spell this out in greater detail, we can distinguish two ways in which a programme might be Aggregate Population Effective but not in the *ex ante* interests of each participant. First, people may differ in values. Imagine a population where each individual has exactly the same probability of "being helped" or "being harmed" by a screening programme, but individuals differ in the values they place on those outcomes. Insofar as individual preferences play some role in determining their well-being, and by extension their prospects, the same programme might be in the *ex ante* interests of some (those who really want to avoid breast cancer and care little about overtreatment) but not others (those who are highly averse to unnecessary medical care). Second, people may differ in their "risk." Imagine a population where each individual has identical preferences, but different chances of "being helped" or "being harmed". Suppose, for example, we know that women between 50 and 70 who breastfed are at lower risk of developing breast cancer than women who did not. A screening programme inviting *all* women between 50 and 70 might not be in the *ex-ante* interests of women who breastfed but in the interests of those who did. Notice that in both scenarios, it is possible that the programmes were both Aggregate Population Effective. But this does not license us to conclude that the programmes were thereby in the *ex-ante* interests of each affected individual.

So, if the UK National Screening Council's ninth criterion is a way of expressing Aggregate Population Effectiveness, then this does not directly get at the question of ultimate importance, namely, whether a screening programme is ethically permissible. If, on the other hand, the UK NSC's ninth criterion is a way of expressing *ex ante* DNH, then the principle is far more demanding than is normally recognized.

What about situations in which *ex ante* DNH is satisfied? Suppose Prevention Paradox is one such case. Both 'The High Risk Strategy' and 'The Population Strategy' are in the *ex ante* interests of all affected. In John's terminology, this leads to a 'relative prevention paradox,' a situation in which the policy choice presents a *prima facie* tension between *Save the Greatest Number* and *Concentration of Risk*. The tension is *prima facie* because, strictly speaking, the "all else equal" clause does not hold in *Save the Greatest Number*. After all, 'The High Risk Strategy' and 'The Population Strategy' policies affect individuals at different risks of cancer, so there is a contextual factor in the background that is not constant. Nonetheless, when choosing between policies in

Prevention Paradox, the point is that the two moral principles tug in different directions. In my terminology, this presents a tension between Aggregate Population Effectiveness and *Ex Ante* Individual Effectiveness. Which moral principle should be privileged? Which way of thinking about effectiveness should be privileged?

My point here is simply that whichever way of thinking about effectiveness you prefer, you will be implicitly privileging a certain moral perspective. So, for example, if you think that Aggregate Population Effectiveness is right, then sometimes you must deny *Concentration of Risk.* And this can come at a certain theoretical cost. Consider an example from Frances Kamm. Intuitively, it seems morally problematic to hire two construction workers to build a bridge, when each face a 50% risk of dying in the building process. To go ahead in this situation looks akin to flipping a coin to decide which of two individuals will die, all for the sake of building a new bridge. This looks deeply problematic. But what if, instead of hiring only two workers, ten thousand workers were hired to build the bridge, with each of these individuals facing a 1/10,000 risk of dying. One might still find it worrisome to go head with the building project, but intuitively speaking, it seems that the latter situation is at least easier to justify than the former—even though the expected outcomes are the same in both (one life lost). So, it seems to matter, ethically speaking, when risk is spread across more individuals. Aggregate Population Effectiveness, in denying *Concentration of Risk*, is not well-equipped to explain these moral intricacies.

On the other hand, if you think *Ex Ante* Individual Effectiveness is right, then sometimes you must deny *Save the Greatest Number*. This, too, can come at a theoretical cost. Consider the following two examples, adapted from John (2014):

> Small Risk Differences
> You, the policymaker, are tasked with choosing between two different screening policies. The policies are cost-neutral but only one or the other can be implemented. Here are your options:
>
> 'The High Risk Strategy': target a population of 100,000 people at a 0.5 ten-year risk of developing life-threatening cancer

> ‘The Population Strategy’: target a population of 1 million people at a 0.45 ten-year risk of developing life-threatening cancer

In these cases, there is a small risk difference between the the individuals targeted by the different policies, but a very large difference in the population size. It follows, from *Ex Ante* Individual Effectiveness, that ‘The High Risk Strategy’ should be chosen. This is the policy that improves the prospects of the affected individuals the most. But to accept this involves denying *Save the Greatest Number* in a dubious manner. It is not just that ‘The Population Strategy’ will save a *few* extra lives; it will save a *sizeable* number more. So, rejecting *Save the Greatest Number* here involves bringing about a state of affairs which is significantly worse than the alternative. Now consider:

> Large Risk Differences
> You, the policymaker, are tasked with choosing between two different screening policies. The policies are cost-neutral but only one or the other can be implemented. Here are your options:
>
> ‘The High Risk Strategy’: target a population of 100,000 people at a 0.5 ten-year risk of developing life-threatening cancer
>
> ‘The Population Strategy’: target a population of 100 million people at a 0.01 ten-year risk of developing life-threatening cancer

In these cases, even though there is a large risk difference between the individuals, the difference between the size of the population is massive. Again it follows, from *Ex Ante* Individual Effectiveness, that ‘The High Risk Strategy’ should be chosen. This is the policy that improves the prospects of the affected individuals the most. But to accept this involves denying *Save the Greatest Number* in a dubious manner. Again to reject *Save the Greatest Number* involves bringing about a state of affairs which is significantly worse than the alternative.

We are now in a better position to see why effectiveness constitutes what Alexandrova (2018) calls a “mixed claim.” There are at least two ways claims about effectiveness presuppose normative variables. First, there are a set of broadly “prudential” value judgments underlying notions of effectiveness. Aggregate Population Effectiveness relies on a claim about what constitutes the best

states of affairs. Adjudicating this question, in Population Screening, requires a certain type of value judgment about how to commensurate benefits and harms in a population in order to arrive at some all-things-considered judgment that screening, in fact, leads to better aggregate outcomes. Likewise, *Ex Ante* Individual Effectiveness turns on the notion of prospects, and prospects themselves depend on how individuals value different outcomes. Whether screening improves Anne's prospects depends on how Anne disvalues a false-positive result or overdiagnosis in relation to the chance of having her life saved. Further, it also depends on how Anne is "classified" for the purposes of assigning probabilities to different outcomes. As Chapter 4 will argue, this also has a normative dimension—in brief, the idea is that because there is sometimes a "reasonable range" of probabilities one might assign to different outcomes, a non-epistemic value judgment is required to decide how to mesh uncertain probabilities with the calculation of individual prospects.

Second, there are a set of more straightforwardly "ethical" value judgments underlying notions of effectiveness. This is the issue of balancing two different, attractive moral principles: *Concentration of Risk* and *Save the Greatest Number*. Because Aggregate Population Effectiveness denies *Concentration of Risk*, it lacks the resources to draw a distinction between Population Screening and GEV, between policies with equivalent expected outcomes that spread risk across more or less individuals. Yet because *Ex Ante* Individual Effectiveness sometimes denies *Save the Greatest Number*, it lacks the resources to handle cases like Large Risk Differences and Small Risk Differences above. So, in light of these complexities, it seems unlikely that there is only *one* concept of effectiveness, be it Aggregate Population or *Ex Ante* Individual Effectiveness, insofar as it would be a mistake to presume that *Concentration of Risk* always overrides *Save the Greatest Number*, or *vice versa*. Each notion of effectiveness captures an important class of *pro tanto* moral reasons to understand screening effectiveness one way or the other.

3.5 Toward a Pluralist Account of Effectiveness

To conclude this chapter, I want to make the case for a *pluralist* account of effectiveness, the thought that there are multiple concepts of effectiveness that are all valid ways to think about effectiveness in medicine. As I see it, there are two routes to this conclusion.

The first route follows naturally from the discussion of the previous section. One way of understanding the tension between *Concentration of Risk* and *Save the Greatest Number* is that both principles identify *pro tanto* moral reasons to think in terms of Aggregate Population Effectiveness or *Ex Ante* Individual Effectiveness. Yet while these reasons have moral valence, neither can lay claim to always tracking *the* best way to thinking about screening effectiveness, all things considered. On the one hand, Aggregate Population Effectiveness cannot always be the right approach, all things considered, because sometimes screening policies that are Aggregate Population Effective violate *ex ante* DNH. In these circumstances, the screening policy would be *prima facie* ethically impermissible. On the other hand, *Ex Ante* Individual Effectiveness cannot always be the right approach all things considered, either, because sometimes it entails that the "effective" screening policy is the alternative that saves a significantly fewer number of lives. A pluralist account of effectiveness has the advantage here of carving out the conceptual space for *both* ways of thinking to be right, depending on the circumstances.

The second route to pluralism derives from a particular way of doing conceptual analysis. David Chalmers (2011) argues that one way to avoid "merely verbal disputes" is to identify what theoretical role a given concept is intended to play, and then to develop the concept according to what best plays this role. He writes: "instead of asking 'What is X?,' one should focus on the roles one wants X to play and see what can play that role" (Chalmers 2011, 538). Likewise, instead of asking 'What is screening effectiveness?' a better approach asks what roles one wants a concept of screening effectiveness to play, and to theorize about what would enable the concept to best play those roles.

One upshot of this approach to conceptual analysis is that it enables us to retain the theoretical progress made by the philosophers, mentioned in the introduction, that have developed their own accounts of medical effectiveness. So, for example, Stegenga (2015) argues that an effective intervention is one that "intervenes on causes or symptoms of disease to improve health" (35). His account emphasizes that, to be effective, an intervention must intervene on a genuine disease. Now, one might be worried about what Stegenga's account of effectiveness entails for screening, because some programmes, like cervical cancer screening, do not (directly) target genuine disease

or symptoms and hence may be *a priori* ruled out ineffective—even if they are yielding massive benefits with minimal harms. But this would be too quick. Stegenga is interested in "therapeutic interventions that are intended for treating disease with the end of improving health" (34), and sets aside screening. So, the role he wants a concept of effectiveness to play is very different from the role a concept of effectiveness should play for screening.

Accordingly, this approach by Chalmers (2011) sidesteps—wisely, in my opinion—the issue of what is *really* the concept of effectiveness. I agree with Chalmers that not much hangs on this "residual verbal question." By analogy, there is an intractable and tiresome debate over the concept of disease. "What is a disease?" many philosophers have asked. Broadly, three camps of thought have formed. Roughly, naturalists argue for a value-free account disease according to which certain biological functions are operating below typical efficiency (Boorse 1977). Roughly, normativists argue for a value-laden account according to which disease is "a bad thing to have [or be in], that is such that we consider the afflicted person to have been unlucky" (Cooper 2002). Roughly, hybridists combine these two aforementioned views, claiming that disease involves biological dysfunction which is also harmful. Which of these is *the true* concept of disease? I, for one, am not on the edge of my seat, waiting for the verdict from philosophers.

I used to think, instead, that Marc Ereshefsky (2009) was right in advancing an eliminativist view about the disease debate. Rather than bicker over whether a particular condition is disease *proper*, Ereshefsky (2009) suggested that we should look at the issue differently, that "we should frame medical discussions in terms of state descriptions and normative claims" – that is, in terms of physiological states and value judgements about those states. And once we view the issue from this perspective, we glean enough about the details, about what matters for medical purposes, such that a concept of disease becomes superfluous.

Yet while I think Ereshefsky is on the right track, we need not follow his argument to the final destination. We need not completely *eliminate* the concept of disease. There is no need to proclaim: "I don't care about what a disease actually is; just look at the physiology and consider the value judgements and get on with the medical care!" The spirit of Ereshefsky's approach, following Chalmers (2011), can instead be put less provocatively but more powerfully as: "I don't care about

what a disease actually is; what I care about is the associated role that I want a concept of disease to play. Naturalism may play a crucial role for pathologists for these reasons such-and-such; hybridism may play a crucial role for physicians for these reasons such-and-such; normativism may play a crucial role for screen-detected cancers for these reasons such-and-such, and so on." And on this picture we can retain a concept of disease, without descending into merely verbal disputes. This is good news, in one sense, because like it or not, economic and social consequences often hang on disease classifications (Cooper 2002). Eliminating the concept of disease altogether, therefore, may not be a wise move.

So, what role should a concept of screening effectiveness play? The strategy here has been that we want a concept of effectiveness that plays a role in justifying screening policy. This is not a very controversial claim, I hope, but to cover my bases I will bolster and clarify this claim. First, this already appears to be an "ordinary" role when using the term "effectiveness" in standard discourse. When Sir Mike Richards asserts that: "there is no doubt that screening programmes save thousands of lives every year. However, as part of implementing the NHS' long term plan, we want to make certain they are as effective as possible," his use of the term "effective" does not solely appear to invoke a factual statement to the effect of "screening saves thousands of lives every year." There is a slightly stronger and distinct claim that the term "effectiveness" seems intended to buttress, which is lurking in the background: that current screening programmes *are* justifiable health policies, that they are all-things-considered beneficial but not yet optimal.

Second, the notion of effectiveness plays an important role in guiding medical care. More specifically, the notion helps to pick out which interventions are deemed to be in a patient's interests. Here is one way of putting it, using Stegenga's (2015) terminology: the concept of effectiveness underpins "the two principle aims of medical treatment: disease cure and symptom care" (35). So, for example, it makes no sense to claim that drinking orange juice is "effective" at treating cancer unless there is good reason to believe that drinking orange juice, in some sense, aids in disease cure and/or symptom care.

Of course, what is tricky in the case of screening is that, because we are dealing with populations, it is far less straightforward to determine what is in the "interests" of the individuals affected.

Because of the moral asymmetry between intra- and interpersonal justification, because of the ethical difference between prospects and outcomes, and because of the difference between viewing screening from an individual or population-level perspective, the idea of what is in the "interests" of the individuals affected by a programme requires careful attention to the underlying ethical issues. I am sceptical that there is a *single* way to balance these different issues and perspectives, so a pluralist approach to effectiveness that is mindful of these relevant ethical considerations appears to be the best path forward.

It is worth pointing out that my aim in this chapter was not to offer a solution to reconciling aggregative and non-aggregative approaches to ethics. Rather, my aim here was merely to point out that there are different views of effectiveness which stem from different ethical viewpoints. Of course, this raises the interesting question about the possibility of reconciling aggregative and non-aggregative approaches. I will not settle that debate here, but I note that this is an issue some philosophers have attempted to address. For example, Voorhoeve (2014) defends a principle called Aggregate Relevant Claims (ARC), which purports to arbitrate between aggregative and non-aggregative ways of responding to the claims of individuals. Roughly, ARC recommends a form of maximization under a constraint, where the obligation to maximize derives from the aggregative approach and the constraint derives from the nonaggregative approach. More specifically, ARC recommends choosing the alternative that satisfies the greatest sum of strength-weighted, relevant claims. The strategy to choose the option that satisfies the "greatest sum" of claims stems from the aggregative approach. And the "strength-weight" and "relevance" of claims is unpacked with resources from the nonaggregative approach. An individual's claim is *stronger* the more her well-being would be increased by being helped, and the lower the initial baseline of well-being from which this increase would occur. A claim is *relevant* if and only if it is adequately strong relative to the strongest competing claim. Whether an individual's claim is *relevant*, therefore, constrains when it is permissible to maximize.

An example will be illustrative. Many people think that we ought to save a large number of individuals from being permanently bedridden rather than save one life. Yet, many people also think that we ought to save one life rather than prevent many individuals from suffering a very minor harm, no matter how many individuals could be helped. The ARC, according to Voorhoeve

(2014), can explain these judgments. In the first case, the claims of those who would be permanently bedridden are *relevant* to the claim of the individual whose life we might save, because these potentially bedridden individuals have a sufficiently strong claim relative to the life-threatened individual. Because their claims are relevant, we ought to benefit these potentially bedridden people, provided that there are enough of them. By contrast, in the second case, the claims of those who would suffer a very minor harm have *irrelevant* claims when compared to the claim of the individual whose life we might save. For these individuals, their claims are not sufficiently strong compared to the life-threatened individual because there is very little at stake for each of them. Hence, it would be morally wrong to allow their aggregate claims to outcompete the claim of the life-threatened individual. What is more, because their claims are irrelevant, no number of such individuals, regardless of how large that number might be, could ever have their collective claims outweigh the claim of the single life-threatened individual.

Voorhoeve's (2014) approach is certainly appealing, in the sense that it synthesizes both aggregative and non-aggregative approaches. However, it is worth noting that the view is controversial (Halstead 2016; Tomlin 2017). Halstead (2016) denies that Voorhoeve adequately justifies the nonaggregative component of the ARC, arguing instead that full-blooded aggregation is the right view. Tomlin (2017) critiques the way in which the ARC relies on the notion of *relevance*, arguing that it is either ambiguous or subject to counterexamples. Needless to say, the ARC has not resolved the fundamental controversy between aggregative and nonaggregative approaches to ethics. In light of this, Voorhoeve's arguments in favor of the ARC do not necessarily cut against the pluralism defended here. Indeed, even Voorhoeve seems to hint at a pluralistic view: "ARC embodies *one way* of partially accommodating and arbitrating between these aggregative and nonaggregative approaches" (Voorhoeve 2014, 70, emphasis added). My aims here, however, were more modest. I have not attempted to resolve this fundamental debate in moral philosophy. Rather, my aim was merely to show that notions of effectiveness implicitly rely on positions in these ethical debates.

To conclude this section, I will elaborate on the scope of this pluralism about effectiveness. In this chapter, I have sketched only two interpretations of effectiveness, Aggregate Population Effectiveness and *Ex Ante* Individual Effectiveness. One might wonder whether these are

exhaustive of the reasonable options for notions of effectiveness. I suggest that they are not. It is conceivable, for instance, to think that a screening programme solely aimed at reducing health inequalities is also "effective" in a sense. However, in defense of focusing on the two interpretations of effectiveness articulated in this chapter, I will point out that there is some leeway in "specifying" these interpretations when evaluating screening effectiveness (Richardson 1990).

Take Aggregate Population Effectiveness, which says that all else equal, an effective screening programme is one that leads to the best (health-related) population outcomes. One important aspect to note is that this formulation is neutral on the issue of which consequences matter—for example, it is silent on whether consequences should be assessed using QALYs or DALYs, or lives saved, or whether equality of outcomes should be taken as a positive consequence (Nord 1999). Of course, this is not to say that any consequence could be taken to be relevant for evaluating the effectiveness of screening. It would be odd, for instance, to say that a programme that increases pain medication sales is "effective." Rather, the point is that Aggregate Population Effectiveness grants a decent amount of leeway in adjudicating which consequences should be valued positively, whilst also capturing the spirit of aggregative approaches to ethics. Further, this way of thinking about effectiveness differs fundamentally from the approach embodied by Ex Ante Individual Effectiveness, which I argued implicitly relied on nonaggregative approaches to ethics.

One might wonder, though, whether this leeway entails a worry with the *ex ante* DNH principle defended in the previous chapter. The worry goes as follows. If we should be pluralists about effectiveness, then does this not imply that we should also be pluralists about the goals of screening? And if this is the case, then couldn't someone formulate these goals in a way that didn't require compliance with the *ex ante* DNH principle? While this is a tempting thought, it seems to me that *ex ante* DNH should have moral valence regardless of the goals of screening. In defense of this idea, recall that I argued *ex ante* DNH derives from more fundamental ethical asymmetries related to dong/allowing harm and the "separateness of persons." Yet, just because a screening programme might have different goals does not mean that these fundamental ethical distinctions lose water. The justification for *ex ante* DNH does not rely on premises directly related to the goals of screening. While it may be possible to formulate the goals of screening in a way that did not respect *ex ante* DNH—say, if one wanted to assert that the only goal of screening is to maximize

overall population health outcomes—the independent plausibility of *ex ante* DNH entails that this way of formulating the goals of screening would come at an ethical cost. One might think that, in certain situations, this is a price worth paying. But this is very different from thinking that *ex ante* DNH would be an ethically hollow principle if the goals of screening were different.

A final issue worth elaborating on is the following: what are the circumstances under which Aggregate Population Effectiveness should be the guiding principle, and conversely, what are the circumstances under which *Ex Ante* Individual Effectiveness should be the guiding principle? I do not have a robust answer to this question. I cannot, for instance, advance a general framework articulating the circumstances under which one principle should be taken as primary and different circumstances under which the other principle should take precedence. But nor do I consider it my role to do so. I would suggest that this is not a drawback *per se* of the arguments above. The arguments here aim to inform our judgments about screening, by unearthing the hidden ethical considerations underlying notions of screening effectiveness, rather than definitively resolving the circumstances under which policymakers ought to maximize overall outcomes versus assist those at greatest risk of harm.

Nonetheless, I acknowledge that the arguments here will ultimately turn on a broader account of how to think about moral reasoning when principles conflict. One possible solution comes from Richardson's (1990) proposal to resolve concrete ethical problems by specifying norms. On this model, the central task of bringing norms to bear on individual cases involves not just applying and balancing norms, but rather specifying the norms to be appropriate for the specific context. Hence, when the duty to maximize health-related population outcomes comes in conflict with the duty to save those at greatest risk of serious harm, the strategy to resolve this conflict should be to tailor our norms to the case at hand. For example, if a screening policy affects a particularly vulnerable population, then the abstract notion of health-related population outcomes may be "specified" to a narrower version that understands outcomes in terms of reducing inequalities. Of course, taken alone, Richardson's strategy does not settle the issue of how exactly to think about effectiveness, but it does focus attention on ensuring that the interpretation is apt for the target population for a policy.

3.6 Conclusion

This chapter has argued that the notion of screening effectiveness is far more value-laden than commonly assumed, and that effectiveness exemplifies what Alexandrova (2018) calls a "mixed claim". I clarified two different construals of screening effectiveness, one in terms of Aggregate Population Outcomes and the other in terms of *ex ante* individual prospects. And I argued that underpinning these construals are commitments to different independently attractive moral principles. Unearthing these subtle moral commitments makes it transparent that determining when a screening policy is "effective" requires a sensitivity to not just prudential, but also ethical considerations. The chapter concluded by suggesting a shift toward a pluralist account of effectiveness. In the next chapter, the arguments set forth in Chapter 2 and 3 are extended to situations of uncertainty.

CHAPTER 4

Uncertainty

4.1 Introduction

Suppose you are precisely 60% confident in the claim "Cambridge will win the boat race" and I offer you an even bet on it—if you are right, I pay you a tenner and if you are wrong, you pay me a tenner. Should you take this bet? According to orthodox decision theory, the answer is yes. Accepting the bet improves your prospects, in the sense that your expected earnings are positive. A 60% chance at winning ten pounds outweighs a 40% chance at losing ten pounds. I take it this is fairly straightforward.

But suppose the situation is slightly different. Suppose, instead, that I offer you an even bet on the claim "Cambridge will win the boat race," but at best your confidence in the claim is *imprecise*. Suppose at best your confidence that Cambridge will win is between 10-60%. Now a puzzle arises. Which probability estimate, within this range, should you use to calculate your prospects? Should you take the upper-bound precisification (60%), or the mid-point (35%), or the lower-bound precisification (10%) when calculating prospects? Are all probability estimates within this range permissible to use? How should these imprecise probability estimates be "folded" into a theory of prospects? These questions are important to consider. How you answer them can be a matter of great practical importance. In the bet I am offering you, the answer to these questions can dictate whether accepting the wager improves your prospects (e.g. if you think the upper-bound of 60% should be used) or lowers your prospects (e.g. if you think the lower bound of 10% should be used).

Notice that this puzzle can arise even if we follow the method of calculating prospects developed in Chapter 2. Recall that for the purposes of normative assessment, I argued prospects should be calculated relative to the evidence which it would be reasonable to expect a policymaker to collect

and use. Recall that whether some body-of-evidence is "reasonable" is a function both of epistemic concerns (e.g. whether there is a large effect associated with a risk factor) and social/political/moral concerns (e.g. whether it would be reasonable to expect the state to collect this sort of information). Sometimes, though, even if policymakers collect and consider a reasonable body-of-evidence, this evidence may not straightforwardly suggest a precise risk assignment.

A natural example comes from estimates of overdiagnosis. It seems sensible to think that the independent review of NHS breast cancer screening considered a reasonable body-of-evidence (Marmot et al. 2013). Yet even this review did not definitively settle the question of overdiagnosis rates. The panel notes: "The issue for the UK screening programmes is the magnitude of overdiagnosis in women who have been in a screening programme from age 50 to 70, then followed for the rest of their lives. There are no data to answer this question directly. Any estimate will therefore be, at best, provisional." And the report continues: "unfortunately, although there is agreement on the concept of overdiagnosis, there has been a wide divergence of views on how to estimate the amount of overdiagnosis, with the result that estimates of the frequency of overdiagnosis vary widely, from ~0-50%" (Marmot *et al.* 2013, 2217). So, a reasonable body-of-evidence does not necessarily support precise assignments of risk (e.g. there is a 20% chance that your cancer is overdiagnosed). Sometimes, at best, the reasonable body-of-evidence supports *imprecise* assignments (e.g. there is between a 0% and 50% chance that your cancer is overdiagnosed).

So, the puzzle remains: how should this "uncertainty" be incorporated into a theory of prospects? Let me be very clear, from the outset, about how I will understand "uncertainty" because the term can be ambiguous. One form of uncertainty is "statistical"—for example, the 95% confidence interval around estimates of overdiagnosis may be 0-50%. A different and more colloquial form of uncertainty refers to a lack of knowledge about outcomes—for example, if I undergo screening, I do not know whether I will be benefitted or harmed.[30] The uncertainty I have in mind here concerns the sense most famously explained by John Maynard Keynes:

[30] See Kavka (1990) for a more fine-grained taxonomy of uncertainty.

> By "uncertain" knowledge, let me explain, I do not mean merely to distinguish what is known for certain from what is only probable. The game of roulette is not subject, in this sense, to uncertainty; (...). Even the weather is only moderately uncertain. The sense in which I am using the term is that in which the prospect of a European war is uncertain, or the price of copper (...). About these matters there is no scientific basis on which to form any calculable probability whatever. We simply do not know (Keynes 1937, 213).

So, the sense of uncertainty here involves a situation in which, when calculating prospects, the probability of a relevant outcome is unknown. Of course, you might think that the situation with overdiagnosis is different. To say that the probability of overdiagnosis is between 0-50% is not to say that the chance of overdiagnosis is *completely* unknown. We can be very confident, for instance, that the chance a screen-detected cancer is overdiagnosed is not 90%. However, I will understand "uncertain" situations in contrast to risky situations, in which precise probabilities can be assigned to all relevant potential outcomes. And I will follow the account of prospects developed previously in understanding "risk" and "uncertainty" as subjective notions—that is, I will understand these notions as relating to the epistemic beliefs concerning the chances of relevant outcomes that derive from the prior beliefs and evidence available to the relevant decision-maker. So on this picture, contexts of "uncertainty" occur when decision-makers, after consulting a "reasonable" body-of-evidence, are not in a position to assign precise probabilities to relevant outcomes.

The puzzle of how a theory of prospects should cope with uncertainty is important to resolve. After all, on a theory of prospects, one needs to know not only the value of different outcomes, but also the probabilities an outcome will occur. So, the *ex ante* DNH principle developed in Chapter 2 gives out at the point where we are not in a position to assign precise probabilities to outcomes. Likewise, any account of effectiveness relying on prospects needs a strategy to cope with situations in which we are not in a position to assign precise probabilities. To tackle these sorts of cases, we need to develop some tools that can extend the *ex ante* perspective to situations of uncertainty. The aim of this chapter is to develop one satisfactory solution to this puzzle.

The argument of this chapter unfolds in two parts. In the first half of this chapter, the claim that there is uncertainty around overdiagnosis estimates is established more fully. Section 4.2 introduces the "standard model of medical prediction" (Fuller and Flores 2015), and argues that the assumptions required to relate population frequencies to individual prospects often do not hold in the case of overdiagnosis. This underwrites an argument against the view that decision-makers can assign precise probabilities to an individual's chances of overdiagnosis. Section 4.3 bolsters this conclusion by arguing that, because cancer is an evolutionary process, a mechanistic model of prediction will not help much. In the second half of this chapter, the implications of these imprecise probability assignments are explored for a theory of prospects in Section 4.5.1, for the principle of non-maleficence in Section 4.5.2, for the concept of effectiveness in Section 4.5.3. and for the argument from inductive risk in Section 4.5.4.

4.2 The Risk-Generalization Model

Fuller and Flores (2015) argue that, in the twentieth century, a new model of medical prediction emerged. No longer are health professionals relying only on inference from theory or induction from experience, which were the primary methods of medical prediction beforehand. Evidence-based medicine (EBM) yielded a new era, in which a hierarchy ranks the "quality" of evidence depending on the study method. Systematic reviews, meta-analyses, and randomized controlled trials reign supreme at the top of this hierarchy. Below that are non-randomized studies, such as cohort and case-control studies, and at the bottom are case series and expert opinion. Accordingly, EBM privileges population studies over mechanistic reasoning or induction from clinical experience (Howick 2011). This emphasis on the results of population studies, spurred by the development of clinical epidemiology and an increasing focus on disease prevention in twentieth century medicine, led to the ascendance of a "new standard model of prediction."

This model involves *risk measures*. The basic idea is intuitive enough, but will be worth introducing carefully before proceeding. The epidemiological term for the frequency of outcome O in a population is the *absolute risk* (AR). Formally:

$$AR = (\text{number of O in population}) / (\text{n of population})$$

This absolute risk estimate tells us something important. Imagine trials examining the effect of screening on mortality find that, with screening, the 10-year absolute risk of dying of breast cancer for 50 year-old English women is 1%. Call this the screened absolute risk (ARX). With this risk measure, a natural thought is that we can use the frequency of this outcome in a study population to predict the future frequency of the outcome in a target population. So, for example, if ARX is 1%, then it seems natural to make the following prediction: with screening, in the population of 50 year-old English women, 1% will die from breast cancer over the next 10 years. Recall from Chapter 3 that one reason the absolute risk measure can underwrite a prediction claim is the Law of Large Numbers. Since screening programmes typically affect millions of people, we can predict with high certainty that the number of women who will actually suffer a breast cancer death, under the screening programme, is 50,000 in a screened population of 5 million individuals.

Notice, however, that this information alone tells us nothing about whether screening actually reduces breast cancer mortality. We also need to know the unscreened absolute risk (AR-X): what is the 10-year risk of dying from breast cancer for 50-year-old English women in the *absence* of screening? Contrast, for example, the situations in which the AR-X is 3% as opposed to 0.5%. If the unscreened absolute risk is 3%, then screening reduces mortality risk; we would then predict that screening prevents breast cancer-related deaths. But if instead the AR-X is 0.5%, then screening increases mortality risk; this suggests that screening either increases breast cancer-related deaths or deaths from other causes. The takeaway here is that, in order to get a sense of the overall effect of an intervention, we need to compare both ARX and AR-X. This enables us to calculate the *absolute risk reduction* (ARR):

$$ARR = AR_{\text{-}X} - AR_X$$

So, for example, if in the absence of screening breast cancer mortality AR-X is 3% and with screening ARX is 1%, then the ARR would be 2%. In turn, it seems natural to make the following prediction: implementing screening for 50 year-old English women will reduce the frequency of

cancer mortality by 2%.[31]. In a screened population of 5 million women, this translates into 100,000 lives saved.

We can already begin to appreciate the basic contours of the "model of medical prediction" that Fuller and Flores (2015) have in mind. Their important insight, however, is that this model of medical prediction actually involves two steps, or two stages of "sub-inferences." We saw only the first step above. The first step involves a risk measure in the study population, such as rates of mortality reduction, being *generalized* to the target population. It involves, as we saw above, using the ARR from a study population to make a prediction about the target population.

But to stop here, to be satisfied with only a *population-level* prediction would be lacking for medical practice. Medicine treats individuals. Population-level predictions are only indirectly relevant for individuals. To bridge this gap between populations and individuals, a risk measure needs to be *particularized* to a specific individual from the target population. This second step of medical prediction is "subtle but not trivial; it signals a shift in meaning from 'risk' as frequency of the outcome to 'risk' as probability of the outcome" (Fuller and Flores 2015, 56). For example, imagine Jane, a 50-year-old English woman, is curious about *her* 10-year risk of developing breast cancer. Imagine that Jane is a heavy smoker, and that there are strong links between smoking and increased breast cancer risk. Even if we know the screened absolute risk for 50-year-old woman, it may not be appropriate to equate Jane's risk with the population frequency if, for example, the trial only involved non-smoking woman.

This two-step model of prediction is called the Risk Generalization-Particularization Model, or Risk GP for short. Here is the complete inference scheme, where AR refers to the absolute risk, and r the frequency of outcome *O*:

[31] One might also present the *relative risk reduction*: ARx / AR-x. This is common when reporting trial results for publication. But it can be misleading (Sprenger and Stegenga 2017). I avoid using it here.

In the study population, AR = r for *O*

…..

In target population *F*, AR = r

…..

For patient *a* in target population *F*, p (*O*|*F*) = r

The scheme above illustrates the move from a study population risk measure to a target population risk measure, and then to the probability of an individual developing the outcome. For example, there are studies that measure the absolute risk of a screen-detected cancer being overdiagnosed. Suppose in one study population, the AR of a screen-detected cancer being overdiagnosed was 20% (the top line of the scheme). This AR in the study population can be used to predict the absolute risk of screen-detected overdiagnosis in a target population (the middle line of the scheme). And, finally, an individual probability of screen-detected overdiagnosis may be inferred for a particular patient from the target population (the final line of the scheme).

I take it the basic idea of the Risk GP model is clear enough. EBM led to the emphasis of epidemiological studies measuring the frequency of outcomes in populations. These population frequencies from studies can be used to underwrite a prediction about a target population (generalization). In turn, the individual probability of a patient can be inferred from these population studies (particularization). Though the basic idea of the model may be straightforward, we need to be clear about the tasks at hand. One task is to describe the structure or logic of medical predictions. This was the aim of the present section. Quite a different issue is articulating what this model of medical prediction *presumes* in order to be successful. It turns out that both stages of the Risk GP model rely on assumptions that warrant scrutiny; these assumptions cannot simply be presumed to hold true. In the following section, I will describe the assumption underwriting the generalization step, and argue that in the case of overdiagnosis it often does not hold.

******

Generalization involves the move from the frequency of the outcome measured in a study population to a prediction of the frequency of the outcome in the target population. The crucial

assumption for successful generalization concerns *representativeness*: the study population must be "sufficiently similar" to the target population. This assumption is necessary in order to justify expecting a similar effect (Fuller and Flores 2015).

Here is a natural way of unpacking *representativeness*. Recall from Chapter 1 that Bleyer and Welch (2012) analysed the incidence of breast cancer stages in the United States and found that, after the advent of mammography screening, the incidence of early-stage breast cancer increased and the incidence of late-stage disease only marginally decreased. Recall that this implies that mammography is leading to overdiagnosis, for in the absence of overdiagnosis there should be a concomitant decrease in late-stage cancer incidence. Bleyer and Welch (2012) conclude that this "suggests that there is substantial overdiagnosis, accounting for nearly a third of all newly diagnosed breast cancers" (1998).

Can we assume that "nearly a third of all newly diagnosed breast cancers," in the NHS, are also instances of overdiagnosis? Not necessarily. The study population involved U.S. women, but our focus is on NHS screening programmes. This means our target population is U.K. women. The question we need to ask is whether *representativeness* is warranted. On the one hand, Bleyer and Welch's (2012) study seems to provide clear evidence that mammography screening can and very likely does lead to overdiagnosis in the U.S. On the other hand, though, it is not immediately clearing what bearing their results have for a population of U.K. women. There are salient differences between the United States and the United Kingdom—different healthcare systems, different background rates of cancer, and different ways that screening is carried out. In the United States, screening is "recommended" by institutions, and some of these institutions "recommend" screening beginning at age 40. In the United Kingdom, screening is organized by a central NHS programme, which "invites" women to "make an informed choice" (Forbes et al. 2014) about screening every three years, beginning at age 50. These considerations put pressure on the tenability of the *representativeness* assumption. So, it would be a mistake to take the relevance of Bleyer and Welch's conclusions for the NHS breast cancer screening programme at face value. Rather, careful reflection is necessary to justify thinking that the study results will generalize to the target populations.

These issues are familiar ones in the philosophy of science. Many philosophers have discussed what Fuller and Flores call the *representativeness* assumption, but under slightly different guises. Broadbent (2013) considers the problems raised by *representativeness* in a discussion of extrapolation, arguing that extrapolation hinges on whether the contexts are sufficiently similar. Likewise, Steel argues that "similarity in all relevant respects may be required for extrapolating an exact, quantitative causal effect" from a study population to a target population (Steel 2008, 80). And Nancy Cartwright (2007, 2010, 2011, 2012), in a series of highly influential papers, discusses the problem in terms of external validity: when are we justified in exporting a conclusion from the test population to the target population?

Does the *representativeness* assumption hold in moving from studies measuring overdiagnosis rates to the target U.K. population? Support for *representativeness* comes in two forms. One is methodological. For instance, a study population formed via a large random sample of the target population provides some support for *representativeness*. The study methodology may also need to simulate the right conditions in the target population. For example, the U.K. screening programme invites women triennially over twenty years. *Representativeness* therefore garners more support from studies examining overdiagnosis rates in women triennially screened over twenty years than rates in women annually screened for ten years.

The trouble, for estimates of overdiagnosis, is that these forms of methodological support for *representativeness* are either absent or scant. The study population is usually not sampled from the target U.K population. As the independent review of NHS breast cancer screening noted: "the issue for the UK screening programmes is the magnitude of overdiagnosis in women who have been in a screening programme from age 50-70, then followed for the rest of their lives. There are no data to answer this question." (Marmot *et al.* 2013, 2216). We saw in Chapter 1 that overdiagnosis can be estimated by RCTs, or observational studies, or autopsy studies. While RCTs have the best study design to estimate overdiagnosis, all of the available RCTs were conducted decades ago and examined the impact of screening for periods shorter than those of the NHS breast cancer screening programme. Moreover, the most reliable "catch-up" RCTs estimating overdiagnosis were conducted in Canada and Sweden (Marmot *et al.* 2013). And while historical and autopsy studies

can also estimate overdiagnosis rates, the available studies often draw from populations outside the U.K.

The other type of support for *representativeness* concerns similarities in the causal structures of the study and target populations. Cartwright (2011), for example, develops sufficient conditions for when *representativeness* is warranted, given that the probability of an effect is fixed by its causes. When extrapolating from a study population X to target population Θ for an outcome O:

a) X and Θ are the same with respect to "[t]he causal laws affecting O"
b) "[e]ach 'causally homogeneous' subclass has the same probability in Θ as in X"

To explain, condition (b) posits that the causal variables are distributed in exactly the same way in populations X and Θ, and condition (a) posits that *if* the causal variables are distributed in exactly the same way in populations X and Θ, then it must also be the case that these causal variables contribute to outcome O in the same way. So, for Bleyer and Welch's study results to transport to the NHS context, it must be the case that, say, background rates of breast cancer are distributed equally between the U.S. and U.K. populations and, moreover, that these rates contribute to overdiagnosis rates equally. But this is highly dubious. We know that overdiagnosis rates increase with more frequent screening (Welch and Black 2010). Testing more asymptomatic individuals uncovers more harmless cancers. And we know that screening is often recommended more frequently in the U.S. So, even if we assumed (implausibly) that the background rates of breast cancer were distributed identically between the U.S. and U.K. populations, it follows that these rates would still unequally contribute to overdiagnosis rates given different screening frequencies. Indeed, as Cartwright notes, these sufficient conditions are very epistemically demanding. And Steel (2008) concurs, arguing that since background causes will always differ between separate populations, presuming that a causal claim will extrapolate to a new context is often an unwise bet.

The takeaway, at this point, is that the key assumption necessary for generalizing study estimates of overdiagnosis is on very shaky grounds. But matters are even worse. Consider another serious complication that arises in applying the Risk GP model to estimates of overdiagnosis: even if the

representative assumption did rest on stable foundations, there is a problem that arises prior to generalization. The issue is one that is internal to the first line of the Risk GP model above, to determining the frequency of outcomes in the study population. As Marmot *et al.* (2013) note, "there has been a wide divergence of views on how to estimate the amount of overdiagnosis, with the result that estimates of the frequency of overdiagnosis vary widely, from ~ 0-50%" (2217). In their discussion, Fuller and Flores (2015) presume that measuring the frequency of the outcome in a study population is relatively straightforward, that risk measures in the study population will converge on roughly the same frequency. While this idealization permitted them to articulate their Risk GP model, it is important to note that this in practice there can sometimes be a wide range of plausible frequencies in the study population.

Combining this issue of unclear population-level frequencies with the worries about the *representativeness* assumption above implies that decision-makers are not in a position to assign a precise probability to the chance that an individual's screen-detected cancer has been overdiagnosed. This is the key claim I want to advance in the first half of this chapter, so it pays to spell out the argument here in more detail. The problem begins with the imprecision around study estimates of overdiagnosis. If population estimates of overdiagnosis in study populations are imprecise, then this imprecision seems to "transfer" through to each step in the Risk GP model.

Recall that the key assumption for generalization is *representativeness*. I suggested above that this assumption is on shaky grounds. Even if I am wrong, though, this assumption justifies an expectation of similar effects between the study and target populations—that there would be similar rates of overdiagnosis if the causal structures of the populations are adequately alike. But there is no reason to think that the representative assumption would *reduce* the uncertainty around overdiagnosis rates when moving from the study population to the target population. If anything, our uncertainty should increase, precisely because it is unclear whether the representative assumption is warranted.

This "uncertainty" seems to impact the particularization step, too. Here is an analogy: suppose I have an urn of one hundred coloured balls. I want to determine my probability of drawing a red one. Suppose I am in a context of risk; I know there are exactly 30 red balls. It follows that my

chances of drawing a red ball from the urn is 30/100. Now suppose, instead, that I am in a context of uncertainty; I do not know precisely how many red balls are in the urn. I only know that the frequency is somewhere between 10 and 50. What are my odds of drawing a red ball? That depends on how one deals with situations of uncertainty. The important point, though, is that this "population uncertainty" about the number of red balls in the urn seems to affect the "individual uncertainty" around my odds of drawing a red ball. I can no longer confidently claim that my odds of drawing a red ball are exactly 30/100 (c.f. Ove Hansson 2006).

4.3 The Mechanistic Model

So much for population studies. However, when calculating prospects, it may be possible to gather evidence from a different source and avoid the imprecision of population frequencies. Fuller and Flores (2015), toward the end of their paper, advance a case for *model pluralism*: the view that different models of prediction may be suitable depending on the situation. The Risk GP Model may not carry us very far; its assumptions may not hold. But that does not mean that probability estimates of overdiagnosis are necessarily imprecise. Perhaps we simply need to appeal to a different model of prediction, one relying on *mechanisms* instead (Machamer, Darden, and Craver 2000; Russo and Williamson 2007; Clarke *et al.* 2014).

On this model of prediction, the starting point is a known causal mechanism, say, from *X* to *O.* A prediction is then made that, if *X* is intervened on in this way, then *X* will produce *O* (Andersen 2012). As Fuller and Flores (2015) note, this mechanistic model requires its own host of assumptions: "that the mechanisms are understood in their full complexity, that intermediate components of the mechanisms are intact, that the mechanisms produce *O* with a high enough probability, and that the effect of *O*-producing mechanisms is not masked by the presence of *O*-inhibiting mechanisms" (58). But this approach has an advantage, in that it does not depend on aggregate outcomes in populations. So, the thought is that if we can use the mechanisms underlying carcinogenesis in service of underwriting prediction claims, then it may be able to sidestep the worries raised for the Risk GP model in the previous section. Moreover, mechanistic information such as the genomic sequence of a particular tumour may be "reasonable evidence" to collect. In support of this idea, consider one explicit goal of the recently published NHS long-term plan: "We

will extend the use of molecular diagnostics and, over the next ten years, the NHS will routinely offer genomic testing to all people with cancer for whom it would be of clinical benefit".[32]

Nonetheless, I am not persuaded that appealing to mechanisms will resolve the issue of uncertainty. In what follows I will present a "Darwinian dilemma" for a mechanistic model of prediction. Briefly, the point is that because cancer is an evolutionary process, there will be an element of contingency involved in cancer progression. This contingency limits the epistemic ability to predict how a particular tumour will progress and, I will suggest, the ability to assign precise estimates of overdiagnosis.[33]

******

How is it that normal cells transform into cancer cells able to proliferate uncontrollably and resist cell death? One common answer is that there are certain molecular events—genetic mutations, epigenetic changes—that confer the "hallmark" capabilities of cancer cells (Hanahan and Weinberg 2011). The "hallmarks of cancer" are biological capabilities acquired during the development of cancer. These include sustained proliferative signalling, growth suppressor evasion, resistance of cell death, replicative immortality, angiogenesis, activation of invasion and metastasis, reprogramming of energy metabolism, and evasion of the immune system. Consider the avoidance of apoptosis, or programmed cell death. We might wonder which genes or signalling pathways underwrite this capability. Though there are many ways to realize it, the most common is by a loss-of-function mutation in TP53, a tumour suppressor gene. This, in turn, removes the damage sensor that induces apoptosis in normally functioning cells. Underlying this answer is the idea that molecular events—usually mutated genes or networks of them—are the primary causal events that beget cancer (Marcum 2005).

[32] https://www.longtermplan.nhs.uk/

[33] I should note a philosophical dispute lurking in the background here. Some philosophers claim that natural selection *is* a mechanism (Barros 2008), in the sense originally articulated by Machamer, Darden, and Craver (2000). Others claim that natural selection is not a mechanism in this sense (Skipper and Millstein 2005). My argument here does not turn on this debate. All I claim is that evolutionary processes, including but not limited to natural selection, may be *characterized* as mechanisms in a model of prediction.

There is another way to answer our question. Instead, the acquisition of the hallmark capabilities of cancer cells can "be rationalized by the need of incipient cancer cells to acquire the traits that enable them to become tumorigenic and ultimately malignant" (Hanahan and Weinberg 2011, 646).[34] The implicit idea here, first described over forty years ago, is that "cancer can be viewed as the operation of Darwinian selection among competing populations of dividing cells" (Nowell 1976; Cairns 1975). Tumours are not simply composed of genetically similar conglomerates of dividing cells; tumour cells exhibit marked genetic and phenotypic diversity. Such variation is heritable across generations of cells and causally leads to differential reproductive success. Hence, tumours are populations of evolving cells, analogous to the evolution of asexual microorganisms. Carcinogenesis is an evolutionary process.

This Darwinian answer is, of course, compatible with the first answer, insofar as we are pluralists about explanations. Explanatory pluralism contends that, for any event, there may be different true narratives that each explain why the event occurred. We need not commit to *either* the molecular explanation or the evolutionary explanation, thinking that only one is *the* true explanation for carcinogenesis (Plutynski 2018). Both can be correct. Some explanations cite more proximate causes, others more distal; some describe the causes at the macro-level, others at the micro-level. In my view, explanatory pluralism is right. There is no objective, mind-independent reason to think one explanation of carcinogenesis—the molecular *or* evolutionary—is the best. Rather, some explanations may be more useful than others, depending on our aims.

And given our clinical aims, there appear to be some substantive reasons to adopt a Darwinian perspective. Consider the following issues that evolutionary theory helps to clarify:

> *Tumour Heterogeneity*. Genomic analyses reveal dramatic heterogeneity among cells within a single tumour. This finding holds across most types of cancers (Campbell et al. 2008). Why is this important? Genetic heterogeneity provides the heritable variation on which selection can act. And when there are more genetic variants in a population of cells, it is more likely that some variants will have proliferative or survival advantages in terms of fitness. This implies that increased tumour heterogeneity should be associated

[34] The language here sounds very teleological. There is a debate, within the philosophy of biology, concerning whether natural selection is "the only important cause of" the evolution of a trait (Orzack and Forber 2017). The subsequent analysis is neutral on this issue.

with a more aggressive progression of disease, which is indeed observed in breast cancer (Park *et al.* 2010).

Implications of Heterogeneity for Treatment. Why do tumours often recur more aggressively after treatment? Because the heterogeneity of tumour cells increases the likelihood that some tumour cells possess resistance to therapy, and cancer treatment clears the ecological niche for such resistant cells to expand—often more aggressively (Aktipis and Nesse 2013). Here is an analogy: spraying fields with pesticides sometimes leads to selection for resistant pests, since the chemicals may not eradicate all of the pests. Once the pesticides destroy most of the pests, however, those remaining resistant ones are not only more prevalent, but they also have more ecological space for expansion.

The Role of Tumour Microenvironment. How is it that there is a surprisingly high frequency of cancer-causing mutations in our normal cells, and yet the vast majority of us do not have cancer? For example, Martincorena *et al.* (2015) found that sun-exposed skin cells contain thousands of somatic point mutations, a similar number to many cancers, yet such cells still maintain the normal physiological functions of the skin. One plausible answer involves the interaction between evolutionary units and their environments. Biologists have recently begun to appreciate the role that the microenvironment plays in tumour progression (Bissell and Hines 2011). In brief: the microenvironment provides crucial signalling to maintain tissue architecture, inhibit cell growth, and suppress malignant phenotypes. But it is a double-edged sword: incorrect signals can destabilize tissue homeostasis and promote the transformation of normal cells to malignancy.

These evolutionary considerations have implications for overdiagnosis. First, the heterogeneity of cancer cuts deep. So deep, in fact, that no two tumours, even of the same cancer type, are thought to share an identical genetic profile (Stratton, Campbell, and Futreal 2009). Intra-tumour heterogeneity entails inter-tumour heterogeneity—the differences within a single tumour imply that there are (notable) differences in "disease" between individuals with the same "diagnosis". Yet the molecular profile of cells within a tumour is one key factor that determines the clinical course of a cancer. Evolution, in this sense, explains overdiagnosis. The relatively crude diagnostic criteria for cancer does not track the underlying molecular differences that shape the evolutionary progression of cancer i.e. whether the tumour progresses to become harmful or not.

Second, one "prerequisite" for overdiagnosis is the existence of a disease reservoir (Welch and Black 2010). In other words, there must actually be a number of symptomless cancers which are detectable but not worth detecting. After all, screening does not cause the incidence of harmless disease to increase; it uncovers the disease reservoir such that the *recorded* incidence of disease goes up. For example, in the case of prostate cancer, it is estimated that in men older than 60 there is a reservoir of detectable disease in the range of 30-70% (Damiano *et al*. 2007). Disease reservoirs consist of cellular abnormalities that meet the pathological criteria for cancer but never develop to cause symptoms. Bissell and Hines (2011) note that "it is evident that at least indolent or occult tumors occur much more frequently than is commonly recognized but are restrained from progressing into overt cancer;" they gesture at the microenvironment as one process involved in this suppression. Hence, the role of the tumour microenvironment in cancer suppression provides one explanation for the gap between "pathological" cellular growths and "clinically significant" cancers.

Third, even if we acknowledge the inter- and intratumour heterogeneity of cancer, it is not easily addressed in clinical practice. Because a single biopsy does not capture the heterogeneity of other sections of a tumour, there is generally a practical constraint on the ability to characterize a cancer, insofar as there are other active molecular pathways that affect tumour cell growth not captured by the biopsy snapshot (Tannock and Hickman 2016). Here is an analogy: imagine a dish full of genetically heterogeneous bacteria culture. Suppose we are interested in whether this population of bacteria will develop the capacity to survive temperatures above 100 degrees Celsius. Were we to assay this population, sequence the genomes of our sample, analyse the activated cell signalling pathways, and so on, it is not at all clear that we would have a good grip on whether the population in the dish will evolve the heat-resistance trait of interest. An assay is a synchronic and incomplete snapshot. But understanding (and predicting) the complexity of evolutionary phenomenon may require a diachronic perspective, with a fuller appreciation of factors not captured in that snapshot (Okasha 2006).

Stepping back from the details for a moment, notice that, for screen-detected cancers discovered in otherwise healthy individuals, the distinction between harmful and harmless cancers implicitly relies on a claim about the future trajectory of that tumour. Almost all cancers, when detected by

screening, are by stipulation harmless, insofar as they are asymptomatic.[35] So the issue relevant to overdiagnosis primarily concerns the future progression of the disease. The task of distinguishing between harmful and harmless cancers then becomes a task of predicting whether, and how, a tumour will progress. Yet we have some knowledge of how this occurs: tumours progress via the somatic evolution of the cell lineages that comprise them. Overdiagnosis, therefore, is an evolutionary issue.

This is some progress, for it provides a different way of thinking about prediction and overdiagnosis. But just how far will it take us? I have a pessimistic attitude. My pessimism stems from the problem of evolutionary contingency. Let me explain.

Philosophers working on evolution make a distinction between evolutionary *products* and evolutionary *processes* (McConwell 2017). Setting aside some niceties, evolutionary products are the outcomes of evolution. Some examples are the Krebs cycle for metabolism, or the human eye, or the cellular ability to invade and metastasize to other tissues. Evolutionary processes are the "mechanisms" responsible for such products. Some examples include natural selection, mutation, and random drift. To claim that each screen-detected cancer has a 20% chance of not progressing encodes a claim about the chance an evolutionary outcome will occur.

But there is a complication. The basic processes that influence evolutionary outcomes are mutation generation (this provides the heritable variation upon which selection can act), genetic drift (changes in trait frequency due to random events), and natural selection (changes in trait frequency due to differential fitness advantages). As evolutionary theorists have pointed out, however, these evolutionary processes are themselves "chancy processes," in that the same initial conditions, when subject to the influence of mutations and/or drift and/or selection, do not necessarily lead to the same outcome (Beatty 2006; c.f. Gould 1989; Beatty 1995). This has been extensively discussed in the literature around evolutionary contingency, where philosophers are interested in whether certain biological forms are robustly realized across a number of evolutionary scenarios

[35] I set aside the conceptual complexities of defining "harm," whether some "risks" count as harm, and so on (c.f. Lewens 2007).

(Wong 2019), or whether replaying the "tape of life" would have resulted in markedly different biological forms (McConwell 2017; McConwell and Currie 2016; Powell 2009).

The evolutionary contingency literature is vast and complex. But we do not need to get into the details. All we need to note is, first, that the initial conditions for evolutionary change will differ across individuals and, second, the processes that drive evolutionary change from those conditions are stochastic. That initial conditions will differ follows from the problem of tumour heterogeneity. Screen-detected cancers differ in terms of (epi)genetic aberrations. Even supposing that evolutionary processes were "deterministic" (i.e. they produce the same outcomes from the same initial conditions), the difference in initial conditions implies different progressions of disease. Conversely, suppose that all screen-detected cancers shared identical (epi)genetic profiles. Even if this were true, the processes influencing evolutionary change are stochastic, such as the timing and exact genomic location of mutations. It follows that (epi)genetically identical cancers will evolve different evolutionary outcomes and, by extension, progress along different pathways.

In sum: cancer progression—and by extension, overdiagnosis—is an evolutionary issue. But evolutionary processes are influenced by stochastic factors, and this means that even if screen-detected cancers were sequenced and this "mechanistic" information incorporated when calculating risk of overdiagnosis, the evidence still may not straightforwardly suggest a precise number. This is, of course, an empirical claim. But I am not alone in expressing worries that the evolutionary nature of cancer imposes "limits" to our ability to predict cancer progression (Lipinski *et al.* 2016; Tannock and Hickman 2016). If these commentators are right, then a mechanistic model of prediction—given our current understanding of cancer—does not resolve the problem of uncertainty.

4.4 Taking Stock

Thus far, I have described the Risk GP model of prediction, which Fuller and Flores (2015) argue is the standard model of prediction. I have discussed the key assumption underlying generalization, and argued that it rests on tenuous foundations. Most studies estimating overdiagnosis rates concern populations that may not be *representative* of the target population of English women.

And even if these study populations were sufficiently similar to the target population, these population studies do not suggest a precise frequency of overdiagnosis in the population. Equally, I have suggested that appealing to a different model of prediction, one relying on mechanisms instead, will not be much help. Because overdiagnosis is an evolutionary phenomenon, and because we currently lack the tools to predict evolutionary outcomes with accuracy, a mechanistic model of prediction will not eliminate uncertainty.

These considerations entail that policymakers evaluating screening will often be in a situation of uncertainty. In other words, the first half of this chapter aimed to establish the claim that screening policymaking *must* deal with situations of uncertainty. In the second half of this chapter, I will explore one strategy for dealing with this uncertainty. Recall that, in order to be satisfactory, the theory of prospects developed in previous chapters needs to be extended to situations of uncertainty. For example, since *ex ante* DNH from Chapter 2 is intended to operationalize concerns around the harms of screening, like overdiagnosis, we need to resolve how the principle should cope with contexts in which the probabilities of harm are imprecise.

One possible answer is to accept a certain amount of leeway when calculating prospects in the wake of uncertainty. To give this idea some motivation, consider that in situations with imprecise probabilities, philosophers tend to say that there is an expansion of what is rationally permissible to choose (Elga 2010; Hare 2010). Think of it this way: if you are 30% confident in a claim and I offer you an even bet on it (if you win, I pay you a tenner and if you lose, you pay me a tenner), then you rationally ought to refuse the bet. But if you are between 10-60% confident in a claim and I offer you an even bet, then you rationally might refuse the bet, because this would be the case if you were precisely 30% confident. Equally, you might also rationally accept the bet, because this would be the case if you were precisely 60% confident. Another possible answer would be to take a stance on where, in a boundaried interval of probability, we should place our confidence (Hare 2017). So, for example, if my confidence that Jane's screen-detected cancer has between a 10-50% of being overdiagnosed, then I might find some reason to favour using the upper-bound precisification (50%), or the mid-point (30%), or the lower-bound precisification (10%).

In the remainder of this chapter, I want to discuss a third approach, developed recently by Rowe and Voorhoeve (2019), that clarifies the relationship between uncertainty and prospects. After introducing this approach, I will develop the implications of this approach for the principle of non-maleficence, for effectiveness, and for the relationship between uncertainty and prospects. I will describe this approach in the following section, before exploring its implications.

4.5 Implications of Uncertainty

4.5.1 Prospects

The uncertainty around individual probability raises a curious issue when thinking about how it relates to a theory of prospects. In the previous chapters, I defined prospects as an individual's expected well-being, calculated by summing the prudential value of each possible outcome multiplied by its chance of occurrence. But this account makes an assumption that may not be warranted: we may not be in a position to assign probabilities to outcomes. It assumed that we were making a choice under *risk*. Yet this is very different from a choice under *uncertainty*, where we lack the knowledge to assign precise probabilities. And these contexts of uncertainty are not uncommon. Determining rates of overdiagnosis in a screening programme, I argued above, is one example, but so is work around the impacts of climate change or the dilemma raised in trying to predict how novel strains of influenza virus will spread (Rowe and Voorhoeve 2019).

The problem to address now is how this uncertainty impinges on a theory of prospects and, more broadly, on the principle of non-maleficence and concept of effectiveness developed in the previous chapters which rely on the *ex-ante* perspective. To begin developing an answer, consider the following case:

Ambiguity

Jane has late-stage cancer. You, the doctor, must choose between two treatments. On the one hand, you can provide Jane with a well-established risky treatment, which based on a large body of evidence suggests that it has a 50% of curing Jane and a 50% of having no effect. Call this option *risky treatment*. On the other hand, you can administer a novel experimental treatment, which will either cure Jane or have no effect. But you have no information on the probabilities of these possible outcomes. You also lack precise prior beliefs about the probability this experimental treatment will be curative. Call this option *uncertain treatment.* Which option, *risky* or *uncertain treatment*, would you provide? Which option *should* you provide?

There is ample evidence that most decision-makers would choose *risky treatment* over *uncertain treatment*. Social scientists, in a broad range of experiments, have demonstrated that many people display "uncertainty" or "ambiguity" aversion when faced with such choices. Contrast, for example, a case in which I will buy you a pint if a fair coin lands heads, and a case in which I will buy you a pint if a foreign coin lands heads, but all you know about the foreign coin is that the probability of it landing heads may be anywhere between 0 and 100% (Trautmann and van de Kuilen 2015). Most people prefer their prospects with the fair coin. Most people, that is, display "ambiguity aversion": they prefer risky prospects over uncertain prospects.

The observation that people display ambiguity aversion is a descriptive claim, but what normative implications does this have for our case above? Rowe and Voorhoeve (2019) have recently argued that it is morally permissible to hold an ambiguity averse attitude and favour the risky treatment. Their central thought is that, since your evidence and prior beliefs are compatible with a wide range of probability assignments, any assignments of a precise probability would be arbitrary in the sense that they go beyond the available evidence. It follows that it is at least reasonable to assign a range of possible probability distributions over the outcomes, so long as they are consistent with your data and prior beliefs. In turn, it seems there is some wiggle room for how one assigns "decision weight" to each possible probability distribution. We should, for instance, assign some decision weight to both the worst and the best possible probability distribution (maybe, in the worst case, it is consistent with your evidence that the experimental treatment provides no chance of cure; maybe, in the best case, it is consistent with your evidence that the experimental treatment

provides a certain cure). But just how much decision weight you assign to each possible distribution is, within a sensible range, up to you.

Ambiguity aversion is a permissible attitude because it is within this sensible range of how decision weight is ascribed to possible probability distributions. For the crucial thing to note is that ambiguity aversion merely involves giving more "decision weight" to the least favourable possible probability distribution than to the more favourable one. It may be helpful to consider an analogy with the more familiar notion "risk aversion." Consider the choice between the following two options:

> Option 1: 85% chance of winning $1000 and a 15% chance of winning nothing
> Option 2: receiving $800 for certain

Many people prefer Option 2 over Option 1 (Tversky and Kahneman 1981). They prefer the guaranteed $800 over the gamble, even though the latter confers a higher expectation of winnings (in brief: 85% * $1000 > 100% * $800). One way to understand the phenomenon of "risk aversion" is to note that people may put more "decision weight" on the worst case scenario: the situation in which they go for the gamble, things do not pan out, and they end up with nothing. Often, this is enough to sway people toward the less risky option. By the lights of standard rational choice theory, this behaviour is irrational. But it may nonetheless be reasonable. This line of thinking may generalize to ambiguity aversion, since it likewise may be reasonable to assign more "decision weight" not to the worst possible outcome, but rather to the worst possible probability distribution consistent with the evidence.

One consequence of this reasoning is that it generates a reason, in Ambiguity above, to prefer the *risky treatment* over *uncertain treatment*. Suppose the most pessimistic assignment of probabilities, consistent with the evidence, is that *uncertain treatment* has a 10% chance of being curative, and the most optimistic assignment, consistent with the evidence, is that it has a 90% chance of being curative. If it is permissible for you, the physician, to be ambiguity averse, then this means you can assign a somewhat greater weight to the most pessimistic assignment of probabilities i.e. you can be more cautious and assign greater decision weight to the possibility

that *uncertain treatment* only has a 10% chance of being curative. It follows, according to Rowe and Voorhoeve (2019), that there is a "depressing effect of uncertainty on the value of individuals' prospects" when people are ambiguity averse (268). What is more, it turns out that reducing the range of uncertainty also reduces the depressing impact on the value of prospects. An uncertain treatment, with between a 25% - 75% chance of being curative, offers better prospects than an uncertain treatment, with between a 10% and 90% chance of being curative.

I am perfectly happy to accept that ambiguity aversion is permissible, that uncertainty has a depressing effect on prospects. Rather than scrutinize the appeal of Rowe and Voorhoeve's argument, in the remainder of this chapter I will explore the implications of their arguments for non-maleficence (Section 4.5.2), for effectiveness (Section 4.5.3), and for what constitutes "reasonable evidence to collect" when calculating prospects (Section 4.5.4).

Of course, you may think I am being too sanguine in accepting Rowe and Voorhoeve's argument. After all, there are a number of routes you might want to take to reject it. Perhaps you want to say that an ambiguity averse attitude is neither permissible nor reasonable; hence, it is implausible to think that uncertainty has a depressing effect on prospects. Alternatively, the jury is still out, amongst decision theorists, concerning whether uncertainty aversion is rational (Bradley 2017). So perhaps even if ambiguity aversion is permissible or reasonable, you fall in the camp of decision theorists who believe it is *irrational* to be ambiguity averse. And perhaps you think, like Caspar Hare (2013), that a promising approach to normative ethics is to combine plausible principles of morality and rationality. You may think, as a result, that a view which is only viable, on pain of irrationality, comes at a high theoretical cost. Perhaps you think this cost comes at too high a price to accept Rowe and Voorhoeve's argument.

These argumentative moves are, of course, available to make. Thus, it pays to clarify the scope of this approach to dealing with uncertainty and to offer some arguments in service of justifying it. First, the claim here is not that ambiguity aversion is the *only* sensible attitude when dealing with uncertain situations. Following Rowe and Voorhoeve (2018), the idea is simply that a moderate degree of ambiguity aversion is sensible, and that this aversion can be offered as a good reason for choosing a risky treatment over an uncertain treatment.

Second, it is worth noting that there is indeed disagreement amongst decision theorists concerning the rationality of ambiguity aversion. To elaborate on the cases discussed above, in contexts of ambiguity, outcomes are uncertain and the probabilities of their occurrence are unknown. In a famous thought experiment, Ellsberg (1961) conjectured that people would prefer choosing from an urn which they know contains 50 red and 50 black balls over an unknown urn containing 100 red and black balls, but in an unknown proportion. This "Ellsberg's Paradox" suggests that people prefer decision situations with known probabilities over situations with unknown probabilities. This behavioral trait has been termed ambiguity aversion, and has been empirically confirmed in many experiments (Camerer and Weber 1992).

More recently, there has been a debate in the literature over whether these Ellsberg choices are rational responses to ambiguity. For example, Al-Najjar and Weinstein (2009) argue that ambiguity aversion cannot be rational, because adopting such an attitude induces violations of other attractive rational principles, such as dominance and time consistency, as well as an aversion to new information. In contrast, Gilboa *et al.* (2009) argue for a more flexible notion of rationality, according to which in uncertain situations "there may not be a perfectly rational choice at all" (288). On this picture, when faced with a situation with uncertain probabilities, like screening, it would be impermissible to arbitrarily assign probabilities and make decisions based on these probabilities. What, then, would be the rational thing to do? Gilboa *et al.* (2009) suggest that this requires trading-off different "ingredients" to rational choice. Some of these "ingredients" pertain to the internal coherence of rational choice theory. For example, is violating dominance a theoretical cost too high to pay? Other "ingredients" pertain to the external coherence of rational choice theory with scientific evidence and reasoning. For example, is arbitrarily positing a probability that goes beyond the scientific evidence a cost too high to pay for a theory of rationality?

The approach I am endorsing here follows Gilboa *et al*. (2009). In situations of uncertainty, there is a tension between different "ingredients" of rationality, and different ways of trading off these "ingredients" are sensible. Further, one sensible way of making these trade-offs involves the use of ambiguity averse decision principles. Here, I am also agreeing with Heal and Millner (2014):

"There is unlikely to be a knockdown argument that determines whether ambiguity aversion is a behavioral anomaly or a rational response to uncertainty, and both positions contain valuable insights into how we might want to make choices under uncertainty" (130). Accordingly, the debate over the rationality of ambiguity aversion is a topic about which reasonable people can reasonably disagree. Insofar as this is the case, it seems worthwhile to explore the relationship between Rowe and Voorhoeve's (2018) approach to dealing with uncertainty and the preceding arguments that relied heavily on a theory of prospects.

To conclude this section, I will address one last worry about Rowe and Voorhoeve's arguments. It goes as follows. Rowe and Voorhoove assume that it is permissible to be ambiguity averse and that, in turn, uncertainty has a "depressing effect" on the value of individuals' prospects. But one might worry that this thought is incompatible with prospect theory as defined by Tversky and Kahneman (1981). After all, Tversky and Kahneman show that whether choices are framed as losses or gains can make a difference to risk attitudes. When individuals are faced with a choice between a guaranteed benefit and a lower chance of a significantly larger benefit, such that the latter option entails higher expected utility, it turns out that people tend to be risk averse in this domain of gains. They opt for the guaranteed benefit, even though this choice entails less expected utility than the alternative. Conversely, when individuals are faced with a choice between a guaranteed loss and a lower chance of a larger loss, such that the latter option entails lower expected utility than the former, it turns out that people tend to be risk seeking. They opt for the lower chance of a larger loss, even though this choice entails less expected utility than the alternative. In light of these framing effects, one might worry that ambiguity aversion could not only decrease, but also increase, the value of an uncertain prospect relative to a certain or risky one, depending on how the alternatives are framed.

I have two responses. First, it is important to note that there may be ways to get around the problem of framing effects. For example, Chwang (2015) discusses the concern that framing effects may invalidate the moral validity of consent when a patient would, say, opt for a treatment when it is described in terms of gains but decline the same treatment when it is described in terms of losses. How can we get around the problem of framing effects here? As Chwang (2015) suggests, what we want is a way to ensure that a patient's "decision is invariant across all frames, so that we do

not have to defend some particular frame as privileged or neutral" (279). When this is the case, we can essentially eliminate the framing effect and "debias" the patient. This strategy may be helpful in the case of screening, too. When policymakers are faced with decision-making in contexts of uncertainty, the evidence may be presented in different ways to ensure that framing effects are not the deciding factor in decisions. While this may not completely eliminate all framing effects, it may at least be helpful in significantly reducing it, such that ambiguity aversion will be consistent across different frames. As such, it seems possible for ambiguity aversion to be a robust enough phenomenon for the purposes of ethical evaluation.

Second, a recent empirical study suggests that ambiguity averse attitudes will be present regardless of whether choices are framed in terms of losses or gains, at least in the health domain. To elaborate, Attema *et al.* (2018) found that there is a difference in ambiguity attitudes between health and money. As noted above, in Tversky and Kahneman's (1981) classic experiments, when options were framed as choices between two losses, people tended to be risk seeking. However, in many of Tversky and Kahneman's examples to illustrate the framing effect, the choices involved money, yet ambiguity and risk attitudes have been shown to be domain specific—the attitudes differ depending on whether the situation involves money or health (Hardisty and Weber 2009). While the literature suggests that people are both risk and ambiguity seeking for losses of money (Baillon and Bleichrodt 2015), there appear to be different attitudes for losses of health. In particular, Attema *et al*. (2018) found that for health losses, people tend to be ambiguity averse instead. This ambiguity averse attitude for health losses suggests that, when choosing between a drug with between a 30-70% chance of fatal side effects, and a drug with a 50% chance of fatal side effects, people tend to choose the drug with the known 50% chance of fatal side effects. Attema *et al.* (2018) found a similar though less strong ambiguity averse attitude for health gains. Hence, when choosing between a drug with a 30-70% chance of gaining 6 months of life and a drug with a 50% chance of gaining 6 months of life, people tend to choose the drug with the known 50% chance of gaining 6 months, though this aversion to ambituity was less strong than with health losses. These empirical findings are noteworthy for the arguments in this chapter, because they suggest that people value reductions in clinical ambiguity, regardless of whether the options are framed as gains or losses (Attema *et al.* 2018, 1713). As such, the worry that ambiguity aversion

could either increase or decrease the value of a prospect depending on the framing of options does not appear to be supported by the empirical evidence.

4.5.2 Non-Maleficence

Heidi Malm (1999) has written a fascinating paper about the inconsistency between screening and non-maleficence. At first blush, her arguments share many affinities with the worries about non-maleficence raised in the previous chapter. Both arguments start from the premise that screening inevitably imposes severe harms on some individuals. Both defend a *prima facie* requirement for ethically justified screening: that "the test reasonably can be expected to be *beneficial on balance* for the person taking it," where Malm (1999) understands the qualifier "on balance" along the lines of prospects (27). And both are sceptical that appealing to consent can help to wriggle out of this problem. For Malm, this is because "it ignores the fact that persons often consent to the test *because* of the recommendations," a worry particularly salient in the U.S. system with institutions encouraging particular groups to get tested (36). For our purposes, the appeal to consent fails because of the well-documented psychological tendencies to reason poorly about cancer risk, and because it ultimately does not address the worry about non-maleficence.

Nonetheless, there is a fundamental difference in how our arguments justify the conflict between screening and non-maleficence. For Malm's predominant worry is the absence of good trial data that constitutes evidence that screening actually does increase prospects. Accordingly, her focus is on "whether physicians and other health care professionals are justified in *encouraging* apparently healthy persons to submit to the tests in the absence of clear evidence that doing so will be good for them, on balance" (35). In essence, then, Malm advances a very different reason for why screening violates non-maleficence: epistemic uncertainty around whether screening is in the expected interests of those invited, stemming from the lack of good trial data. By contrast, the arguments raised in Chapter 2 claimed that non-maleficence is violated if screening is not actually in the expected interests of all. What is more, the arguments from Chapter 2 presumed that policymakers had a firm grasp on the impact of screening policy on prospects.

What to make of Malm's argument? Taken alone, I do not find it very compelling. After all, over the past two decades, there has been more trial data published, and though we are nowhere near consensus on the magnitude of the benefits and harms of screening, we have surely reduced our uncertainty about such considerations. It would follow, if Malm were right, that somehow screening violates non-maleficence *less* than it did twenty years ago. But I find this conclusion rather odd. It is not immediately clear how this reduction in epistemic uncertainty means that screening imposes less morally objectionable harm. It is perfectly consistent, in fact, for our uncertainty about the benefits and harms to decrease and for screening in recent years to have imposed more unnecessary harms.

But Rowe and Voorhoeve's argument puts an interesting twist on Malm's link of thinking. For if uncertainty has a depressing effect on one's prospects given ambiguity aversion, then this may at least partly illuminate how Malm was on to something. She is right that uncertainty matters but wrong about why. It is not that epistemic uncertainty about the benefits and harms of screening poses a *direct* conflict with non-maleficence; rather, that epistemic uncertainty may have a depressing effect on the value of prospects. And this deflation in prospects may be just enough to tip the scale in favour of thinking that screening no longer is in the expected best interests of all. Uncertainty, then, plays a more *indirect* role in creating a tension with non-maleficence than Malm's (1999) argument suggests.

To see this more clearly, suppose policymakers are considering a programme which is very uncertain, in the sense that there are imprecise estimates of the benefits and harms of screening. Suppose, however, that in spite of this uncertainty, the policymakers judge the programme to just *barely* be in the interests of those affected. The programme is implemented at time 1. Then a decade passes, and at time 2, when more trials have been published, we now have more knowledge about the impact of screening on mortality, such that the estimates of the benefits of screening are more precise. And at time 2, more "catch up" studies have been published, so we have more knowledge about the rates of overdiagnosis in this screening programme, such that the estimates of the harms of screening are more precise. If the only difference between time 1 and time 2 is that the probabilities assignments became more precise (i.e. the new trials did not make the probabilities assignments more precise *and* suggest that screening has a much greater impact on

mortality than previously thought), then the basic idea is that the programme at time 2 raises the prospects of those affected *more* than it did at time 1. This is because, at time 1, the uncertainty around the probability estimates decreased the value of the prospects compared with at time 2.

This has an interesting implication for screening policymaking. For if it is the case that uncertainty depresses the value of prospects, then screening programmes with uncertainty probabilities of benefit and/or harm are more difficult to justify implementing than courses of action with more certain probabilities. For example, there is often rather shaky evidence of the impact of population screening, because randomized trials are difficult to conduct as discussed in Chapter 1. But, in contrast, there tends to be good evidence of the impact of treatment on cancer outcomes, such that it is much easier to assign probabilities. Accordingly, when there are treatments for a particular cancer with well-established probabilities of benefits and harms, it may be more difficult to justify implementing a new screening programme that has far more uncertain probabilities of benefits and harms.

4.5.3 Effectiveness

Recall from Chapter 3 that I argued different interpretations of screening effectiveness can pull us in different directions. On the one hand, an interpretation of effectiveness that looks to population outcomes may recommend a "Population Strategy," targeting many individuals at low risk of developing cancer. On the other hand, an interpretation of effectiveness that looks to prospects may recommend a "High Risk Strategy," targeting fewer individuals at greater risk of developing cancer. In this section, I will highlight one implication that arises from Rowe and Voorhoeve's arguments: that the permissibility of uncertainty aversion may wedge a larger gap between these two interpretations of effectiveness than Chapter 3 let on.

Recall that one view of screening effectiveness says that a screening programme is effective if it leads to better population outcomes. I argued that, at least sometimes, a reason to reject this interpretation of effectiveness is that it can lead policymakers to implement screening when it cannot be justified to each individual—i.e. when screening lowers the health-related prospects of some. I suggested that this follows from a plausible principle of *ex ante* DNH: any screening

programme not in the *ex ante* interests of all invited is *prima facie* impermissible. One reason to accept *ex-ante* DNH, I suggested in Chapter 2, was that it can align the *ex ante* and *ex post* perspectives. After all, if a programme is in the expected interests of all, then given the Law of Large Numbers, it will also lead to better overall population outcomes relative to the *status quo.*

However, I also said in Chapter 2 that *ex ante* DNH is silent on the choice between two options that raise the *ex ante* interests of all. So, in Chapter 3, I sketched two different ways of thinking about effectiveness that can help guide the choice between different screening policies, depending on how we think about "effectiveness." I argued that these different accounts of effectiveness can be understood to operationalize different moral principles, for example, pertaining to the concentration of risk. In this section I will develop another route to this conclusion. In brief, the point is that if ambiguity aversion is permissible, then there is another way the *ex ante* and *ex post* perspectives can diverge, in addition to those already discussed in Chapter 3. And this is the case, even when all of the policy choices satisfy *ex ante* DNH. Let me explain. Consider the following choice between Policy 1 and Policy 2 (Rowe and Voorhoeve 2019):

> *Policy 1*: Gives an uncertain chance of (Anne cured, B very ill) in State of the World 1 OR an uncertain chance of (Anne very ill, Betty cured) in State of the World 2, where "uncertain" denotes some unknown probability between 0 and 1
>
> *Policy 2*: Leaves both Anne and Betty equally but slightly ill in States of the World 1 and 2

Suppose the utility of being cured is 80, of being very ill is 50, of being slightly ill is 65 – c, where c is some cost associated with choosing this policy and $c \geq 0$. Suppose that both policies raise the prospects of Anne and Betty, because the *status quo* of doing nothing is that both suffer a painful death. In other words, both policies are permissible by the lights of *ex ante* DNH. Here are the different possible states of the world, in table form:

Action	Person	State of the World 1 ($0 \leq p \leq 1$)	State of the World 2 ($0 \leq p \leq 1$)
Policy 1	Anne	80	50
	Betty	50	80
Policy 2	Anne	65 - c	65 - c
	Betty	65 - c	65 - c

Table 1. Policies Under Uncertainty

Notice that if ambiguity aversion implies a depressing effect on prospects, then this effect only arises in evaluating Policy 1. Under Policy 1, the value of Anne and Betty's prospects are depressed because the decision-maker is unable to assign probabilities to which state of the world will eventuate. On the other hand, Policy 2 is not affected by this uncertainty. Although the chance that a given state of the world will eventuate is unknown, Policy 2 is a scenario in which there is equality of prospects and outcomes. No matter *which* state of the world occurs, Anne and Betty will end up with 65 – c utility. Because individual prospects under Policy 2 are not depressed by the effect of uncertainty, it follows that an ambiguity averse decision-maker should be willing to accept a small cost ($c > 0$) in prospects—at some positive value of c, there will be a threshold at which the prospects for Anne and Betty are equivalent under both Policy 1 and 2.

Suppose that c is positive but small. Suppose that c is so small, in fact, that the depressing effect of uncertainty involved in opting for Policy 1 entails that Policy 2 yields better prospects for both Anne and Betty. It follows, from a view of effectiveness concerned with prospects, that Policy 2 should be chosen. Policy 2 raises the prospects of Anne and Betty more than Policy 1, and recall that this captures a link between an account of permissibility and action guidance: the decision-maker chooses the alternative that can be justified to each person on the grounds it was chosen for her own sake. Recall that, on this picture, a course-of-action is justifiable for a person's sake just in case it increases the prospects of that person. But notice something curious. For *any* situation in which $c > 0$, Policy 1 leads to better overall outcomes, in the sense that the sum total of well-being is higher. For *any* situation in which $c > 0$, it follows, from a view effectiveness concerned with aggregate outcomes, that Policy 1 should be chosen. This is a surprising result! For it means that

sometimes, even when a set of policy is in the *ex ante* interests of all invited, the *ex ante* and *ex post* perspectives on effectiveness can still diverge.

What is the moral here? *Ex ante* DNH carves out a *prima facie* constraint on permissible policies. Even within this space, however, "effectiveness" considerations —depending on whether the *ex ante* or *ex post* perspective is privileged—can still tug in different directions. Or, to put the point slightly differently, "effectiveness" considerations hinge on how much decision weight we put on certain moral perspectives. These perspectives concern the concentration of risk, or aggregate outcomes, or as this section discussed, how much caution we exhibit in the face of uncertain probabilities, which in turn depresses the value of prospects. How we adjudicate these moral issues will impinge on which policy is most "effective" and, more broadly, on which policy should be chosen all-things-considered.

4.5.4 Reasonable Evidence

In this section, I want to make a final observation that will smooth the transition to the final chapter of this thesis. The observation is that, if uncertainty has a depressing effect on prospects, then there is an unnoticed connection here with another classic debate in the philosophy science. The other classic debate concerns the proper role for values in science. I have in mind, in particular, the argument from inductive risk, which purports to show there is an appropriate role for non-epistemic values in scientific justification (Douglas 2000).

Here is a rough sketch of the argument from inductive risk: when is the evidence sufficient for accepting, asserting, or acting on a claim? Roughly speaking, deciding where to set this evidentiary threshold requires attention to the moral consequences of error. So, for example, a higher evidentiary threshold would be required to act on the claim "this drug will not kill you" than the claim "this batch of belt buckles is not defective" (Rudner 1953). Determining the moral consequences of error requires a non-epistemic value judgment. Hence, non-epistemic value judgments indirectly impact the decision of where to set the evidentiary threshold.

Broadly speaking, the argument from inductive risk involves an inference from evidential underdetermination to a discrete set of options, such as accept/reject/suspend judgment (Steele 2012). These options exhaust the logical space of possibility, and the standard approach to deciding which option to choose involves an indirect appeal to non-epistemic values, according to proponents of inductive risk. So, for example, if the error of implementing an "ineffective" screening programme is worse than the error of not implementing an "effective" programme, then the evidentiary threshold for implementing a new programme ought to be somewhat high. This argument will be discussed in far greater detail in Chapter 5.

However, discussions of inductive risk have focused heavily on whether to *accept* or *reject* an inference, with less discussion of the third option of suspending judgment (Kaivanto and Steel 2019). Here I want to suggest that combining ambiguity aversion with the argument from inductive risk suggests two less obvious policy reasons to "suspend judgement and wait for more evidence" when faced with a situation of uncertainty. First, a peculiar upshot of Rowe and Voorhoeve arguments is that, typically, when more evidence comes in, decision-makers are in a better position to assign probabilities to different outcomes. When more evidence comes in, our uncertainty *usually* decreases. Because of this, we can sometimes reasonably expect that the prospects of those affected by a policy under consideration will be higher, after more evidence comes in, than in an evidential state with significant uncertainty—all else being equal. Of course, this *ceteris paribus* clause is doing quite a bit of work here, but set that aside for the sake of argument. If this is right, and if the *ceteris paribus* clause holds, then there may be an additional reason to favour "suspending judgment" or harbouring an "epistemically cautious" attitude in the face of evidential underdetermination: adopting this attitude in response to inductive risk implies that in the future, when there is more evidence and less uncertainty, the prospects of those affected will be higher than simply accepting or rejecting a policy in the face of severe uncertainty.

Second, one curious aspect of screening programmes in particular is that, when deciding whether or not to implement a new policy, there is an asymmetry between the uncertainty of the magnitude of benefits and the magnitude of harms. Even if the magnitude of the benefits of screening are unclear, any screening programme will certainly cause harm—from false-positive results or unnecessary anxiety or time lost. So, typically, there is less uncertainty around the harms of a new

screening programmes than the benefits. As Muir Gray, former Director of the UK National Screening Committee, famously declared: "All screening programmes do harm. Some do good as well and, of these, some do more good than harm at reasonable cost" (Raffle and Gray 2007, xi).

One example of this comes from current discussions over whether to implement an ovarian cancer screening programme. The UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) is an RCT examining whether ovarian cancer screening reduces mortality (Jacobs *et al.* 2016). While there is evidence demonstrating that a blood test can detect ovarian cancers early, we saw in Chapter 1 that this is not the same as early detection making a meaningful difference in health outcomes. Some preliminary results of the UKCTOCS were published in *The Lancet* in 2016. The results suggest that, compared with no screening, a yearly blood test reduces ovarian cancer mortality by 20%. Roughly speaking, this means that for every 10,000 women annually screened with a blood test, around 15 lives could be saved. But the problem is that there was substantial uncertainty around this estimate. Due to the low numbers of women who have developed and died from ovarian cancer in the RCT thus far, the confidence interval was wide— the "true" estimate of mortality reduction is likely to be anywhere between 0% and 40%.

Yet while the benefits of ovarian cancer screening are unclear (there may be no benefit at all or there may be significant benefit), the harms imposed by ovarian cancer screening are certain. For example, any screening programme will inevitably involve false alarms. The UKCTOCS trial suggests that for every three women who undergo diagnostic surgery to check for ovarian cancer, based on the results of the blood test, *two of them* turn out to not have cancer. Unnecessary surgery is a very serious harm. And the fact that it is inevitably imposed on individuals by ovarian cancer screening programme needs to be taken seriously. There is an important moral here, specific to the case of screening: this asymmetry in the uncertainty around the magnitude of benefits and harms of screening programmes implies that the prospects associated with the "uncertain" benefits of a novel programme may be less than commonly presumed.

4.6 Conclusion

This chapter aimed to accomplish three tasks. The first was to advance the claim that many aspects of screening policymaking involve situations of uncertainty i.e. these are scenarios in which there is no scientific basis to assign precise probability assignments to outcomes. The second task was to reframe overdiagnosis as an evolutionary problem. Once overdiagnosis is viewed from a Darwinian perspective, I argued that this raises complications for relating population statistics to *ex ante* prospects on a mechanistic model. Finally, the third task was to explore the relationship between uncertainty and the topics discussed elsewhere in this thesis: non-maleficence (Chapter 2), effectiveness (Chapter 3), and inductive risk (Chapter 5).

Stepping back from the details, the broad structure of the problems raised in this chapter is this. At a certain point, the available evidence "gives out," such that there is a certain amount of leeway in what to believe or do. It may be that your cancer will progress and it may be that your cancer will not progress. The evidence does not conclusively indicate which it will be. Nor, in many cases, does the evidence license a very precise calculation of probabilities concerning which outcomes will occur. But despite this leeway, a choice must be made. A choice must be made about how to "fold" indeterminacy into a theory of prospects, about how to relate this uncertainty to a principle of non-maleficence or to a theory of effectiveness.

With the discussion from section 4.5.4 in mind, the structure of this problem might look eerily familiar. Here is a related but distinct problem. Policymakers often must decide whether to assert a scientific claim about screening, like "mammography screening reduces breast cancer mortality by 20%." Yet it is often the case that the available evidence does not unequivocally vindicate this claim, that the option of waiting for more evidence is not available. In these cases, there is a gap between the evidence and scientific claim, but often a choice must still be made concerning whether to accept or assert or act on this population claim. In these cases, the relevant question becomes this: when is the evidence sufficient for accepting or asserting or acting on a scientific claim? And, as we saw above, according to the "argument from inductive risk," answering this question requires an implicit appeal to non-epistemic values (Douglas 2000). The next chapter will examine the relationship between this argument and screening in more detail.

CHAPTER 5

Underdetermination

5.1 Introduction

The running theme, thus far, has been that the screening debate harbours an underexplored moral dimension that needs to be illuminated and developed. In Chapter 2, we saw that if we take the most plausible interpretation of the non-maleficence principle seriously, then any screening programme which is not in the *ex ante* interests of all affected is ethically impermissible. This is a demanding moral principle, certainly much more demanding than any current screening guideline lets on, and it specifies an ethical constraint to guide screening policy. In Chapter 3, we saw that the concept of "effectiveness" is far more value-laden than commonly presumed. I argued that thinking carefully about different ways to operationalize screening effectiveness leads us to a pluralist account of effectiveness, according to which programmes can be "effective" in distinct ways by relying on different underlying moral principles. Then, in Chapter 4, we grappled with a worry: relating population claims to *ex ante* prospects may be trickier than Chapters 2 and 3 of the thesis intimated. I explored one set of ethical tools that enables us to extend a theory of prospects to situations of uncertainty, drawing on recent work by Rowe and Voorhoeve (2019) concerning the depressing effect that uncertainty may have on the value of prospects.

In this chapter, I want to explore an entirely different kind of argument that purports to carve out a moral dimension in the screening debate. Briefly, the argument gets off the ground when there is a certain type of "uncertainty" around empirical hypotheses. This type of uncertainty is distinct from the "uncertainty" discussed in Chapter 4, so it pays to spell out the difference. In the previous chapter, I understood "uncertainty" as a situation in which decision-makers are not in a position to assign precise probabilities. Sometimes decision-makers are not in a position to proclaim: "the chance a screen-detected cancer is overdiagnosed is 20%." Sometimes, based on the available evidence, they can only forthrightly say: 'the chance a screen-detected cancer is overdiagnosed is

between 10-50%." In these latter situations, Chapter 4 suggested that one way this uncertainty can matter, ethically speaking, is by decreasing the value of a prospect associated with these imprecise estimates, in the sense that it lowers the estimation of the expected utility estimate. All else equal, an imprecise chance of a benefit or harm is less valuable, from the *ex ante* perspective, than a precise chance of a benefit or harm.

Here is a slightly different variant of "uncertainty." For clarity, I will refer to it as "underdetermination." To explain, sometimes the evidence does not empirically or logically entail that the claim "the chance a screen-detected cancer is overdiagnosed is 20%" should be adopted. Sometimes the evidence *underdetermines* theory choice. As Chapters 1 and Chapters 4 discussed, claims about screening are often underdetermined by the evidence. The evidence does not conclusively indicate that *those* are the correct probabilities to report. It may be that screening reduces breast cancer mortality by 50%, or it may be that screening reduces breast cancer mortality by 1%. All of these claims may be compatible with the evidence.

This raises a question: when is the evidence sufficient for accepting or asserting or acting on a particular claim? This chapter will be concerned with how to answer this question. According to the "argument from inductive risk," it is permissible and arguably even desirable to answer these questions by appeal to non-epistemic values (Douglas 2000; 2009). Roughly, the thought is that, because adopting empirical claims can have consequences of moral significance, the evidentiary threshold is or should be sensitive to the "moral consequences of error." So, for example, if regulators are deciding whether to approve a drug for headaches that may increase the risk of kidney failure, then they ought to require a high evidentiary standard. Why? Because if the regulators approve the drug, thinking that it has no side-effects when in fact it does, then the consequences are severe. Many people will suffer kidney failure. But if the regulators do not approve the drug, thinking that it does cause kidney failure when in fact it does not, then the consequences are not so bad. After all, there are other drugs available to alleviate headaches.

The relationship between this "evidential underdetermination" and the "uncertainty around probability estimates" discussed in Chapter 4 is worth spelling out in more detail. On the one hand, both "uncertainty" and "underdetermination" concern situations in which the evidence "gives out,"

so to speak. For example, Rowe and Voorhoeve (2019) write that "uncertain situations" in the sense they have in mind are common, citing the International Panel on Climate Change (IPCC) practice of reporting only probability intervals, "because the best available information does not suffice for the assignment of precise probabilities" (241). The IPCC make statements such as: "It is *likely* [official translation: there is a chance between 0.66 and 1] that land temperatures over Africa will rise faster than the global land average, particularly in the more arid regions."[36] In a similar vein, philosophers writing about "underdetermination" are concerned with situations in which evidential considerations do not definitively point in favour of one hypothesis or the other, such that the question of when the evidence is *sufficiently* strong then turns on an ethical value judgments about the consequences of error.

On the other hand, a clear contrast exists between this "uncertainty around probability estimates" and how philosophers writing about inductive risk understand "evidential underdetermination." The key difference concerns how non-epistemic values are used as tools to cope with "uncertainty" or "underdetermination." Rowe and Voorhoeve (2019) are trying to *reduce* "uncertainty" into an issue of prospects. The line of argument takes uncertainty as a point of departure to underwrite a claim about how one might *directly* value the prospects of different courses-of-action. In turn, because it is permissible to cautiously assign greater decision weight to the worst possible probability distribution than to the better ones, this "uncertainty" can be dealt with by noting that uncertainty has a depressing effect on the value of an individual's prospects.

Commentators on inductive risk, by contrast, use underdetermination to advance the claim *that* values have a legitimate role to play in scientific justification. On this picture, underdetermination carves out a much more *indirect* role for values to play in addressing equivocal evidence. Douglas (2009), for example, is very careful note that indirect roles for values are the only ones acceptable in decisions about whether to accept or reject a hypothesis. She writes, "cognitive, ethical, and

[36] IPCC, "Africa," in Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, ed. Christopher Field, Vincente Barros, David Dokken, Katharine Mach, Michael Mastrandrea, T. Eren Bilir, Monalisa Chatterjee, Kristie Ebi, Yuka Otsuki Estrada, Robert Genova, Betelhem Girma, Eric Kissel, Andrew Levy, Sandy MacCracken, Patricia Mastrandrea, and Leslie White (Cambridge: Cambridge University Press), pp. 1199–243 at p. 1202.

social values all have legitimate *indirect* roles to play in the doing of science, and in the decisions about which empirical claims to make that arise when doing science" (Douglas 2009, 108, emphasis added). Clarifying this point, Douglas explains, "the indirect role for values in science concerns the sufficiency of evidence, the weighing of uncertainty, and the consequences of error, rather than the evaluation of intended consequences or the choices themselves" (Douglas 2009, 103). Philosophers writing about inductive risk, then, are not so much concerned with reducing "underdetermination" into an issue of values, so much as they are trying to use "underdetermination" to understand the relationship *between* values, scientific justification, and the proper place of science in society more broadly.

With the relationship between Chapters 4 and 5 of the thesis clarified, the discussion will unfold as follows. In Section 5.2, I introduce the argument from inductive risk, which purportedly carves out a role for non-epistemic values in science. Following that, in Section 5.3, I discuss an objection to the argument from inductive risk by Betz (2013), which instead asserts that non-epistemic values can be expunged from science with the use of hedged claims. I argue that this objection ultimately fails, and that in the context of screening, there is an appropriate role for non-epistemic values. In Section 5.4, I discuss a recent proposal by John (2018) that advances one way to adjudicate between appropriate and inappropriate influences of non-epistemic values. According to John (2018), the non-epistemic values used in scientific justification should not be incompatible with the values held by the putative audience. Section 5.5 critically examines and extends this reasoning to screening.

Before proceeding, it is worth highlighting that the claim that there is an appropriate role for non-epistemic values in science policy is not new. This is a point that has been commonly advanced in the fields of both science and technology studies (e.g. Owens 2015) as well as history and philosophy of science (e.g. Douglas 2009). The present chapter uses screening as a case-study to explore one specific set of ways in which science policy is value-laden. Further, the chapter critically examines a recent proposal by John (2018) that purports to illuminate *which* non-epistemic values are acceptable in scientific reasoning by appealing to the ethics of communication. The contribution of this chapter, then, is two-fold. First, by closely analyzing examples of screening policymaking, the arguments here illuminate another domain in which

scientific justification for policy is appropriately value-laden. Second, in examining John's (2018) proposal, the chapter develops an emerging body of literature in philosophy of science that treats communication as a crucial concept underpinning the proper role of values in science (Franco 2017; Wilholt 2013). However, while this emerging literature takes communication to be central, the focus is limited to speech acts of *assertion* (Franco 2017). The chapter therefore motivates a shift in thinking more broadly about how other speech acts, such as offers or invitations, may likewise illuminate legitimate roles for values in scientific practice.

5.2 Evidence and Values

5.2.1 *Underdetermination*

This section discusses the argument from inductive risk that leads to the conclusion that non-epistemic values have a proper role to play in scientific justification. The argument gets off the ground by noting the *underdetermination* of theory by evidence. A theory is underdetermined when the evidence does not necessarily imply a particular choice of theory. In these scenarios, there is a gap between logic and evidence, on one end, and theory choice, on the other. Three different versions of underdetermination can be distinguished (Kitcher 2001):

> Transient underdetermination: *some* theories are underdetermined by logic and the *currently available* evidence
> Permanent underdetermination: *some* theories are underdetermined by logic and *all possible* evidence
> Global underdetermination: *all* theories are *permanently* underdetermined

For the evidential situation of screening, there is certainly transient underdetermination. One way of understanding the arguments from Chapter 4, concerning the imprecision of overdiagnosis estimates, is in terms of transient underdetermination. The claim that "a NHS screen-detected cancer has a 20% chance of being overdiagnosed" is not deductively entailed by the evidence. Some evidence, after all, suggests that the chance is much higher. Other evidence suggests that the chance is much lower.

For our purposes, it will not be necessary to advance any claim about permanent underdetermination nor global underdetermination, but I will make one observation about the former variant. One unavoidable aspect of gathering evidence for screening is that it not only requires a great deal of resources, but also *time*. Conducting an RCT to measure the benefits of screening—to determine the impact of screening on mortality reduction—typically takes decades to complete. Conducting a "catch-up" study to measure overdiagnosis rates takes even more time. Recall that at the end of an RCT, it is expected that the number of cancers detected in the screening group will be higher than the control group, simply because screening detects cancers earlier-than-otherwise. In order to quantify overdiagnosis, the control group must be followed for much longer, to see if the number of cancers in this group "catches up" to the number in the screening group. Many things can change in healthcare during that time. Treatments for cancer will predictably improve; diagnostic technologies may become more sensitive. These factors affect the relevance of trials examining the benefits and harms of screening from decades earlier. In Fuller and Flores' (2015) terminology, the results of these trials may no longer be *representative* of the target population. So, in light of these complexities around the "possibility" of gathering good evidence from screening trials, there may be reason to suspect that claims about screening benefits or harms are permanently underdetermined.

Transient underdetermination is a weak claim. There are numerous areas of health and science policy that are transiently underdetermined. Screening is in good company, in this respect. Indeed, Kitcher (2001), writing about transient underdetermination in relation to "objectivity in science," dismisses transient underdetermination as "familiar and unthreatening" (30). Familiar and unthreatening though it may be, transient underdetermination is all that is required to get a certain line of thinking—called the argument from inductive risk—going. This argument is important to explore, because it purports to show that there is a proper role for non-epistemic values in scientific practice. What is more, the argument from inductive risk needs to be clarified, because philosophers in the recent literature have formulated it in different ways. In what follows I will describe this argument, relate it to screening, and provide some examples of inductive risk "gone awry" in the past.

5.2.2 *The Argument from Inductive Risk*

The classic presentation of the argument from inductive risk is found in Rudner's (1953) paper. On this picture, scientists accept or reject hypothesis. Yet the evidence does not always deductively entail whether to accept or reject a given hypothesis. As a result, accepting a hypothesis requires a value judgment about how the risks of reaching a "false-positive" or "false-negative" conclusion should be traded-off against each other. And this value judgment has a moral flavour, because it depends on an evaluative judgment concerning how bad each error is. So, for example, a higher evidential threshold would be required to accept that a drug is not lethal than to accept that a batch of belt buckles is not defective (Rudner 1953). Accordingly, the thought is that "how sure we need to be before we accept a hypothesis will depend on how serious a mistake would be," where the seriousness of mistakes reflects the "moral standards" of those who make them (Rudner 1953, 2) Rudner's argument suggests that "the scientist as scientist does make value judgments" (ibid.).

Heather Douglas (2000, 2009), in recent years, has extended this argument in a variety of ways. Her focus is on the assertion of scientific claims, rather than acceptance. Her emphasis is that scientists have a moral responsibility to consider the consequences of their work, not merely that they do make such value judgments. And she extends the argument from inductive risk to many of the decisions "internal" to science, such as the generation of evidence, the classification of individual data points, and the implicit assumptions underlying statistical analysis of data. In all of these stages of scientific research—stages which are traditionally understood as "purely scientific" matters—Douglas (2000) argues that non-epistemic values play an essential role.

Consider her example of interpreting the carcinogenic effects of dioxins in laboratory rats. There is disagreement, in the scientific community, on the following mutually inconsistent claims:

> *Threshold View*: there is always a threshold for the toxic effects of dioxins; below a certain dosage, dioxins are safe
>
> *No Threshold View*: there is no safe dose of dioxins which does not have carcinogenic effects

Which view is right? Douglas notes that there are arguments for both views. Indeed, "depending on which aspects of the evidence one chooses to emphasize or, more generally, which background assumptions one adopts," both views above are plausible, and "there is significant uncertainty in adopting either" (573). This is an evidential situation of (at least) *transient underdetermination.* Yet adjudicating between the views is necessary. One view or the other must be adopted because a decision must be made about how to regulate dioxins. If the *Threshold View* is adopted, when it is actually wrong, the regulations will likely fall short of protecting public health. Many people will be afflicted prematurely of cancer, and this is a costly moral mistake. If the *No Threshold View* is adopted, when it is actually wrong, the regulations will likely be far too stringent. This may have undesirable economic consequence for certain corporations. So, because the choice between the *Threshold View* and *No Threshold View* has non-epistemic consequences, "non-epistemic values are needed to evaluate the risks of adopting either position and their accompanying background assumptions" (Douglas 2000, 573). Douglas concludes that, in the decision above, the *No Threshold View* should be adopted because the costs of error are comparatively worse.

We have here the gist of an argument that non-epistemic values have a proper role to play in science. Despite differences in how the argument from inductive risk is formulated, the basic outline of the argument from inductive risk occurs in two steps. First, inferences must be made in the move from uncertain evidence to a discrete set of options. These options might include accepting or rejecting or suspending judgment, or they might involve asserting or denying or refraining from comment (Steele 2012). Second, non-epistemic value judgments are relevant in adjudicating this move from evidence to inference. For if we grant that the inference is underdetermined by the evidence, and if we grant that a decision must be made before all of the evidence comes in, then it follows that non-epistemic values *must* play a role. The evidence does not tell us what to infer from the evidential uncertainty; it is the moral and political values that indirectly do (Biddle 2013). To illustrate, consider the following examples.

Here is an editorial, from 1993, in favour of PSA testing for prostate cancer:

> The National Cancer Institute is conducting a prospective, randomized trial to determine whether or not screening reduces the prostate cancer mortality rate, but it will take sixteen years to complete this study. It is estimated that half a million men will die of prostate cancer before this study is completed, and it is unrealistic to expect clinicians to refrain from using PSA for cancer detection in the meantime (Catalona 1993).

And here is another passage, from an advocacy group for early detection:

> It is unconscionable for any agency, public or private, to block lung cancer screening for high risk populations on the basis of a flawed study which will not be completed until 2009 or beyond. During that time, another one million people will die of lung cancer (Lung Cancer Alliance, in Croswell *et al.* 2010).

In both of these cases, the impact of screening on mortality is underdetermined by the evidence. After all, this is *why* the trials are being conducted. Yet, in spite of this underdetermination, the passages advocate the implementation of screening. There is a move from evidential underdetermination to a particular course-of-action. The passages, in other words, exemplify the type of inferences characteristic of inductive risk arguments.

Kaivanto and Steel (2019) note that, in discussions of inductive risk, we can ask at least two distinct questions. One concerns the *evidential threshold*: where exactly is or should the evidential threshold be placed? It should be apparent that, in the cases above, the authors are advocating a rather low evidential threshold. The trials examining screening have not even been *completed* and there is support for implementing the tests. Is this problematic? That depends in part on another question Kaivanto and Steel (2019) point out can be asked about inductive risk cases. The question, in their terminology, involves *reverse engineering*: what values are implicitly embedded in a given inference? If the values that indirectly impact the evidentiary threshold are problematic, then we have reason to suspect the inference is problematic, too. If the values that indirectly impact the evidentiary threshold are acceptable, then we have reason to accept the inference, too.

Asking the *reverse engineering* question puts us in a better position to understand why the passages are problematic. To see why, notice that the views above make an implicit and controversial value

judgment. In particular, the passages assume that the implementation of "false-positive" screening is *costless*. What is important to note here is that this non-epistemic value judgment indirectly impacts the amount of evidence (or lack thereof) required to justify using a screening modality. After all, if we think that offering a screening test has no harms, then this suggests that the evidentiary threshold to accept the effectiveness of the test ought to be low. Why? Well, maybe it is the case that the test really is effective, or maybe it is the case that the test really is not. But if there are no salient harms of screening, then an attitude of "better safe than sorry" seems to be a good approach. Absence of evidence may be evidence of absence, but if we are wrong we shoulder the burden of those thousands of lives lost.[37] That is a costly mistake. And it is certainly worse than offering an ineffective test that we presume to be harmless.

However, once we acknowledge that screening can seriously harm many people, the value judgment that screening is costless becomes problematic. As we saw in the previous chapters, there are many harms of screening and, what is more, these harms are nearly certain to happen. Any screening programme, for instance, will lead to some false-positive test results given that no test is perfectly specific. Moreover, many screening programmes will lead to overdiagnosis and unnecessary medical treatment. If implementing an ineffective screening test is not costless because it imposes serious harms on healthy individuals, then we have good reason to think the extremely low evidentiary threshold advocated in the passages above is problematic.

5.3 An Objection: Betz on Value-Freedom

If we accept the argument from inductive risk, then there appears to be an indirect role for non-epistemic values in science. But the argument is controversial. Critiques of the argument deny one of the two steps that get it off the ground—they either reject that inferences must be made from uncertain evidence to a discrete set of options or they reject that non-epistemic value-judgments play a role in this move. For example, Jeffrey (1956) challenges the first. His thought is that scientists should not settle on a discrete set of options. Instead, scientists should merely report the probabilities of hypotheses given the available evidence. Alternatively, Levi (1960, 1962)

[37] It is worth noting that the motto "absence of evidence isn't evidence of absence" has been discussed in the philosophical literature. See Sober (2009) for an insightful analysis.

challenges the second. His thought is that even if a discrete decision such as accepting or rejecting a hypothesis is an appropriate task for scientists, these decisions should be arbitrated solely by epistemic values such as simplicity or explanatory coherence.

Betz (2013), more recently, also challenges the second move. His thought is that by making uncertainty more transparent in the form of hedged hypotheses, scientists can sidestep making non-epistemic judgments. This section will discuss and address this objection from Betz. I argue that this response to inductive risk will not work because it fails to account for important aspects of scientific communication. If the argument from inductive risk is framed in terms of communication, rather than acceptance, then there is an appropriate role for non-epistemic values in science, even when scientists make nearly certain assertions in the form of hedged hypotheses. Cancer screening provides a good example of this line of thinking, but it can also be extended to mitochondrial donation (Lewens 2019) or neonicotinoid research (John 2014).

One reason to be wary of what the argument from inductive risk purports to show is that it seems to contradict an intuitively appealing account of the scientific process. Call this the *Value-Free Ideal* (VFI): "the justification of scientific findings should not be based on non-epistemic (e.g. moral or political) values" (Betz 2013, 207). Is there anything to be said in support of this view, beyond just insisting that it is intuitive and right? For Betz, at least, the VFI "derives, straightforwardly and independently, from democratic principles and the ideal of personal autonomy" (207). For political decisions that implicitly rely on scientific findings, the VFI ensures "that collective goals are determined by democratically legitimized institutions, and not by a handful of experts" (ibid.). For private decisions, the VFI ensures that the scientific findings on which we rely for routine decision-making are not "soaked with moral assumptions" (ibid.). Accordingly, there are some appealing reasons to uphold the VFI.[38]

Betz (2013) notes that, in order for the argument from inductive risk to have bite, scientists must run the risk of being mistaken. But there is a way to sidestep or seriously mitigate this risk of error. He writes: "Policymaking can be based on hedged hypotheses that make the uncertainties explicit,

[38] Note that another justification for the VFI concerns the desire to avoid wishful thinking in science. I discuss the issue of wishful thinking in more detail below.

and scientific advisors may provide valuable information without inferring plain, unequivocal hypotheses that are not fully backed by the evidence" (212). So, for example, scientists need not claim that overdiagnosis rates in breast cancer screening are exactly 20%. This claim is not deductively entailed by the available evidence, and given the uncertainties around this number sketched in the previous chapter, asserting this claim is prone to inductive risk. Rather, scientists can report a *range* of probabilities which, in turn, makes the uncertainties transparent. Were a scientist to claim, for instance, that overdiagnosis rates in breast cancer screening are between 0-50%, then this "hedged claim" is far more likely to be true given the evidence. Indeed, in these situations, when scientists can be practically certain that a hedged claim is true, the risk of error becomes irrelevant. This suggests that non-epistemic value judgments concerning the consequences of error are not always needed in science.

The basic idea underpinning Betz's argument seems appealing. It is certainly true that the more a claim is hedged, the more certainty one can have in it. But this increase in certainty comes at a cost. The cost that I will focus on concerns how appealing to hedged claims overlooks important aspects of scientific communication which, in turn, undermines the idea that hedged claims can salvage the VFI. For example, Franco (2017) has recently argued that if the argument from inductive risk is framed in terms of *communication*, and more specifically in terms of the speech act of *assertion*, then there are salient and legitimate roles for nonepistemic values in scientific practice. His thought is that scientific practice cannot merely involve individuals *accepting* different hypotheses, for this way of thinking about scientists implicitly assumes that science is primarily concerned with arriving at the correct cognitive attitudes.

Franco is concerned with rebutting a view, by Mitchell (2004), that carves out a distinction between "values appropriate to generating *the belief* and the values appropriate to generating *the action*" (Mitchell 2004, 250). On this picture, scientific practice is mainly in the business of generating correct beliefs or any other relevant cognitive attitudes like acceptance. Thus, for Mitchell (2004), non-epistemic values have a role to play in science, but this role is *external* to scientific practice—theory choice is a purely cognitive matter sensitive to only epistemic criteria, and values enter only after belief in a hypothesis has been settled by the available evidence, for example, when *using* scientific findings to generate policy recommendations. It is one thing to

generate beliefs, which is a cognitive matter, and quite another thing to act on these beliefs, which is a practical matter. On this view, then, the determination of overdiagnosis rates in screening is a purely cognitive matter answering only to the epistemic values appropriate to scientific reasoning. And this is consistent with non-epistemic values adjudicating how those cognitive attitudes inform policy recommendations.

Yet, as Franco (2017) rightly notes, scientific practice does not merely involve individual(s) accepting or rejecting hypotheses; it also involves making *public* the acceptance or rejection of hypotheses, for example, through a journal publication. So, following Longino (1990) who emphasizes the social dimensions of science, Franco highlights that scientific practice is social in the sense that it requires making scientific claims available for scrutiny within a community of researchers. And one way this occurs is through the speech act of *assertion*: "claiming that a hypothesis is (likely) true or false given the available evidence and taking responsibility to demonstrate as much" (Franco 2017, 167). Unlike cognitive attitudes such as acceptance, an assertion communicates one's acceptance of a claim and, in addition, carries with it the implication that others ought to adopt a similar cognitive attitude (Grice 1957). In this way, assertion is clearly a public action undertaken that can potentially have nonepistemic consequences for the audience.

These considerations of scientific communication place us in a better position to see why Betz's position is problematic. Betz notes that proponents of inductive risk typically focus on epistemically risky claims such as "the rate of overdiagnosis in breast cancer screening is 20%." Formulated this way, however, the claim runs a significant chance of turning out to be false. Yet this epistemic risk can be mitigated by making a more hedged claim, such as: "the rate of overdiagnosis in breast cancer screening is between 10-50%." This hedged claim is strongly supported by the available evidence. If scientists can be practically certain of these hedged claims, then there is no need to reflect on the consequences of error before *accepting* them.

But what about before *asserting* them? Two complications can arise here. First, a downside of hedged claims is that, although one can be more certain of their veracity, there is a trade-off in practical guidance. A claim such as, "there is a 20% chance of rain tomorrow in Cambridge" is epistemically riskier than a more hedged claim such as "tomorrow it will either rain or it will not."

Though the latter hedged claim can be advanced with certainty, if my interest is in whether I should pack an umbrella tomorrow, the hedged claim offers little practical guidance to this end. Sometimes a less certain claim may be worth the epistemic risk when the aim is to assert a claim to guide practical action.

Second, and more importantly, one implication of Franco's suggestion to reframe inductive risk in terms of communication is that, speech acts like assertion can have vastly different nonepistemic consequences. And this is true, even when two claims are equally certain. To illustrate, consider the scenario of reporting scientific claims about rates of overdiagnosis to policymakers. At the end of the 10-year Malmö randomized trial, 741 breast cancers were detected in the screening group, and 591 in the control group. A "catch-up" study was reported after 15 years: the initial difference of 150 cancers reduced to 115 cancers between the groups. So 35 cancers "caught up". The remaining 115 cancers count as "overdiagnosis." But here is a residual complexity: how should the estimate of overdiagnosis be reported to policymakers?

On the one hand, we can say that overdiagnosis was 16% (115 in 741), because this was the proportion of cancers in the screening group that were overdiagnosed. On the other hand, we can use a different denominator, because some of the cancers detected in the screening group of the trial were detected in the clinic, and clinically detected cancers with symptoms do not, by some definitions, count as overdiagnosis. Specifically, the trial showed that 64.4% of cancers in the screened group were picked up by screening; the rest were identified clinically. This suggests that 477 of the cancers were screening detected (741*.644). And using this as a denominator, the rate of overdiagnosis is 24% (115 in 477) (Welch and Black 2010). These estimates are equally certain, because they derive from the same study, but the pragmatic effects of these assertions on policymakers may be entirely different. For example, some policymakers may find that the overdiagnosis rate of 24% is far too high, and therefore recommend against screening. Conversely, others may find that the overdiagnosis rate of 16% is acceptable, and therefore recommend in favour of screening.

Of course, one might object that these nuances can simply be relayed to the policymakers, and that doing so would mitigate the different pragmatic effects of these estimates. Accordingly, once these

nuances are communicated as well, then the assertion of these overdiagnosis rates would be "value-free." However, truly satisfying "value-free" communication requires the inclusion of the nuances above and much more: some say that even more follow-up time is needed for the cancers to "catch up," such that the overdiagnosis estimates above are inflated (Hanley 2011). Others say that the above analyses are problematically assuming that background incidence rates are stable over time. They say that there may well have been a "natural increase in incidence" in cancer in the past few years, and this would lead to the overestimation of overdiagnosis (Kopans, Smith, and Duffy 2011).

Should these niceties be included in the communication process to policymakers as well? There appear to be strong arguments in favour of leaving out these nitty-gritty details. After all, policymakers should not be overburdened with information that is not straightforwardly relevant to the decision. At a certain point, the inclusion of all of these scientific subtleties waters down the overall point that the scientist is trying to convey to policymakers—a bit like a physiology lecturer, so keen on explaining the physics of blood flow to medical students, that the clinical relevance of the course is lost on the audience. Equally, however, omitting such niceties for the ends of parsimonious communication rests on the non-epistemic assumption that the audience would prefer more "clean" estimates of overdiagnosis over more "accurate" ones that makes the uncertainties explicit. Thus, omitting such minutiae either concedes that communication is no longer value-free or mitigates the intelligibility and usefulness of the scientific claims being communicated. The point here is that this communicative process is laden with non-epistemic values, because a concern for the practical consequences of scientific assertions should impact how claims are communicated.

To conclude this section, here is one final worry for the VFI in the context of screening. Even if we grant that the VFI is viable and/or desirable, any account of scientific justification that cannot make sense of the actual claims made about screening has a reality problem. Consider the highly contested policy of breast cancer screening for women between the ages of 40-49 in the United States. It is striking that, amongst the seven institutions that publish breast screening guidelines, there is virtually no agreement on screening guidelines for "average risk" women between 40-49.[39]

[39] https://www.cdc.gov/cancer/breast/pdf/BreastCancerScreeningGuidelines.pdf

The U.S. Preventive Services Task Force claims it is an "individual decision;" the American Cancer Society claims it an "individual decision" for women aged 40-44, but those aged 45-49 should get mammograms annually; the American College of Radiology recommends annually screening this age group. Only the International Agency for Research on Cancer explicitly maintains an agnostic attitude, stating that there is insufficient evidence to recommend for or against screening.

These guidelines are communicative acts. They are not merely informing the public about what the institution thinks is in the best interests of women between 40-49; they are also implicitly trying to *persuade* women that they should undergo mammography screening, or to reflect very seriously about whether mammography screening is right for them. Yet, if the same evidence generates such different guidelines for mammography screening, then the inference to the best explanation for why these institutions make inconsistent claims about screening is that such claims are indirectly justified by appeal to different non-epistemic value judgments. It is, for instance, not exactly surprising that the American College of Radiology recommends the most frequent and early start for mammography, given the vested interest in the technology. So, at least in practice, not only do non-epistemic values affect the justification of scientific findings to be communicated, but such values can explain the discrepant claims about screening effectiveness.

5.4 The Value-Apt Ideal

This chapter has argued, thus far, that non-epistemic values can and should play a role in the justification of claims about screening effectiveness. If this broad point is all we can say about the role of non-epistemic values in scientific justification, then none of the examples provided above are necessarily problematic. It is perfectly consistent for the American College of Radiology to assert one guideline, the U.S. Preventive Task Service Force another. They simply endorse different non-epistemic values!

However, this is an unstable position to arrive at. There surely must be some constraints on what sorts of non-epistemic values are acceptable to influence scientific justification. After all, if radiologists are moved to recommend annual screening for women in their 40s for personal

financial gain, then this seems to be an inappropriate influence of non-epistemic values in scientific justification. Wilholt (2009), for example, discusses the phenomenon of preference bias, which occurs when an investigator's preference for a particular result steers the study toward that preferred conclusion over other possible conclusions. Preference bias leads to epistemologically deficient science, according to Wilholt. There is a proper role for non-epistemic values in science, but it does not follow that just any non-epistemic values are permissible.

In the face of these issues, the task is to articulate and justify plausible constraints on the non-epistemic values that can appropriately influence scientific justification. John (2018) has recently developed an instructive line of thinking on this task. This section will critically examine and extend this argument.

John's (2018) idea begins with the following distinction. Sometimes, a scientist engages in *wishful thinking*: "she believes or accepts claims which, given her evidence, are not well-established relative to proper epistemic standards, and this acceptance is motivated by the non-epistemic benefits of accepting those claims, regardless of their truth." And sometimes, a scientist engages in *wishful speaking*: "she makes a claim which, given her evidence, is not well-established, and where her motivation for making that claim is the non-epistemic benefits that follow from others believing (or believing that she believes) that claim, regardless of its truth." These two notions are not always aligned. Sometimes we assert claims that are not well-established for the non-epistemic consequences (saying 'Santa is coming tonight!' to a young child). And sometimes we believe claims that are not well-established, even though we dare not assert them for the non-epistemic consequences (suppose I think everyone will be reincarnated, because I value this comforting thought, but I dare not utter it for fear of ridicule).

Here is another example of wishful speaking. Imagine an epidemiologist who is equivocal, given the evidence, with respect to whether a vaccine has side effects for an unlucky few, but either way she has good reason to assert that the vaccine has no side effects. Maybe the vaccine is beneficial overall for the population, and maybe people will refrain from receiving the vaccine if they think there are side-effects. If so, then she may have reason to assert that the vaccine has no side-effects, even if this is epistemically underdetermined, because this will bring about the best state of affairs

in which the population has herd immunity. Of course, whether the epidemiologist's "wishful speaking" is all-things-considered permissible is a separate matter, but the example nonetheless illustrates that difference between a scientist believing a given claim and asserting a given claim.

The morality of actions may not be exhausted by the consequences of that action, however, at least if we are partial to non-consequentialist views. Regardless of whether asserting a claim brings about good or bad states of affairs, John (2018) points out that wishful speaking is *pro tanto* wrongful because it fails a broadly Kantian principle to respect autonomy: "in speaking wishfully, a speaker treats hearers' beliefs as mere means to be manipulated for the sake of ends which the speaker values. Even if these ends are noble, such an attitude is to disrespect hearers' status as autonomous agents." From this broadly Kantian variant of respect for autonomy, John derives a constraint on non-epistemic values in scientific justification:

> In asserting claims, scientists imply that their audience has reasons to hold those claims. However, if scientists' claims are "based" on non-epistemic values, then audiences who do not share those values may lack good reasons to accept them. Implying that someone should accept a claim based on non-epistemic values which she does not hold is to disrespect her as an autonomous agent. Therefore, communicating claims whose justification rests on non-epistemic value judgments conflicts with the demand that we respect others' autonomy (John 2018, 6).

While this does not necessarily imply that all non-epistemic value judgments in communication are problematic, it does suggest a limit on which values are permissible to import:

> The Value-Apt Ideal (VAI): When we are justifying scientific findings to be communicated to some audience, the justification of those findings should not be based on non-epistemic (e.g. political or moral) values which are incompatible with the values of the putative audience

The VAI is a useful and plausible starting point for thinking through the constraints on values in science. And John's concept of wishful speaking seems to capture the intuition for why "misleading" screening leaflets that only report relative risks and downplay the harms are so morally problematic (Gigerenzer 2015). Such forms of communication not only fail to respect autonomy in the traditional bioethical sense, by hindering understanding and by extension

precluding ample opportunity for informed choice, but also in the additional sense that John develops: there may be non-epistemic benefits that follow from people believing that screening is in their best interests, such as higher uptake (a common metric to appraise the "efficiency" of the screening programme). And if so, then authors of the leaflets seem to be using the audiences' beliefs as a tool to achieve the ends they value.

What I want to highlight here, however, is that John's intended formulation of "wishful speaking" surely does not exhaust the conceptual space of what counts as an "assertion." For example, consider the following two "assertions" that make similar claims but look to be entirely different speech acts. There is a clear difference between Sir Mike Richards publicly claiming (and hence, in a sense, asserting) that screening is effective, and a letter arriving in your mailbox, inviting you to consider the offer of breast cancer screening (Forbes *et al.* 2014). Recall from Chapter 3 that Sir Mike Richards has claimed that: "There is no doubt that screening programmes save thousands of lives every year. However, as part of implementing the NHS's long term plan, we want to make certain they are as effective as possible."[40] This is a very explicit assertion that screening is effective. In contrast, the aim of the NHS screening leaflet is to enable invitees to "consider the offer" of screening (Forbes *et al*. 2014). The idea is that the invitation "seeks neither to encourage screening nor to ask people to make decisions without guidance" (Forbes *et al.* 2014, 195).

Of course, while the screening invitation does not explicitly assert that screening is effective, there is good reason to think that from the perspective of the recipient this claim is implied. For example, reporting of only relative risk reduction on the NHS bowel cancer screening leaflet can plausibly be understood as a "nudge" (Thaler and Sunstein 2008). By communicating risk in relative terms, which has the predictable impact of inflating the perception of benefit from screening (Malenka *et al.* 1993; Hux and Naylor 1995; Sorensen *et al.* 2008), the strategy seems to be to increase the number of people to accept the screening invitation, by implying that screening is effective. However, if the authors of the leaflet do not intend the communication of such risks to imply that screening is effective, then either the reporting of relative risks is not a "nudge," in which case it is simply misleading, or it begs the question of why absolute risk reduction was not reported. In communicating through the screening leaflets, the NHS may want to stay neutral on the issue of

[40] https://www.bbc.co.uk/news/health-46212057

whether screening is effective, yet given the inflated public perception of the benefits of screening (Hoffmann *et al*. 2015) and the privileged relationship to health facts that any healthcare system holds, this seems doubtful in terms of recipient uptake. So, both Sir Mike Richards and the screening offer seem to "assert" that screening is effective, but the way in which these "assertions" are carried out in very different ways. One is very explicit; the other is more implicit but certainly lurking in the background.

Accordingly, if John (2018) is right that the constraint on non-epistemic values in scientific assertion should be derived from norms of communication, then we may need to consider the different types of speech acts that assertions can take. I propose, then, to extend John's argument by distinguishing between his primary focus on wishful speaking in terms of a scientist asserting a claim to a policymaker or to the public, and a rather separate category of speech act in the form of the screening leaflet and accompanying invitation to "consider the offer."

When we invite people to "consider the offer," we open up a new space of medical possibility that needs to be ethically negotiated. And invitations may be governed by certain felicity conditions that will differ from the norms structuring the assertion of scientific claims. Invitations do not impart a neutral set of options nor do they entail a duty to accept. If the screening invitation is doing something with words, and if in terms of uptake this is understood as implicitly asserting that screening is effective, then what are the appropriate felicity conditions? Notice that sometimes the appropriate felicity conditions for an invitation will differ from other types of assertion, even when the non-epistemic values of the putative audience are held constant. For example, when the NHS deploys mobile vans in the parking lot of supermarkets, inviting people to get screened for lung cancer, it is not immediately clear that this invitation is appropriate or ethical, even if the non-epistemic values of the audience warrant the implicit assertion that lung cancer screening is in their interests.

These considerations imply that the compatibility of the audiences' non-epistemic values is not the only dimension that should guide the assertion of scientific claims. We also need to ensure that the assertion is an acceptable category of speech act; and while strictly speaking this is consistent with John's value-apt ideal, there is no reason in principle that the norms of certain speech acts,

like invitations, *must* be compatible with those the putative audience deems fit. For example, even if everyone thinks it is acceptable to be offered lung cancer scans, it does not follow that inviting people to get scanned is appropriate, insofar as the felicity conditions of speech acts may derive from broader, independent social norms. It may be the case that the nature of the interaction, for instance, is incompatible with individuals being able to offer genuine and meaningful informed consent. Were this the case, then even if it might be appropriate to publically assert that lung cancer screening is effective when indirectly using the audience's values, it does not necessarily follow that the speech act of *inviting* people to lung cancer screening is appropriate.

5.5 Mammography Wars

In earlier sections, we saw how non-epistemic values may have a proper role in scientific justification, and how John's (2018) VAI illuminates one potential constraint on which values may be used. But we also saw how a distinction between forms of "wishful speaking" imply differences in the norms guiding assertion, and how this may be independent of the values of the putative audience. In this section, I will explore a potential objection to the VAI, and then highlight the implications for screening policy.

Consider the following historical case-study:

> The Mammography Wars
> In 1996, a consensus conference on breast screening for women aged 40-49 organized by the National Institutes of Health (NIH) was held. Analysis revealed substantial uncertainty about the benefit of screening for this age group: at best, less than one life might be saved if 1000 women are screened for 10 years. But the harms were more certain: at least 250 of those women will have a positive test result and not benefit, many will have unnecessary surgery for inconsequential tumours. The Consensus Panel's verdict, approved by a vote of 10 to 2, was this: "At the present time available data do not warrant single recommendation for mammography for all women in their forties. Each woman should decide for herself whether to undergo mammography" (National Institutes of Health, January 1997).
>
> The response was immediate and furious. The Panel was accused of condemning women to death. The *New York Times* called the report fraudulent. Nightly news

> opened with an apology to women for the report. The Senate voted 98 to zero for a resolution to support mammography. The head of the NIH was shocked by the report and asked the NIH Advisory Board to reexamine the evidence. This is what the NIH Advisory Board did. And they revised the recommendation, issued on March 1997 and approved by a vote of 17 to 1, as follows: "Women in their forties who are at average risk should get a screening mammogram every one to two years" (National Cancer Advisory Board, March 1997). (Raffle and Gray 2007)

The influence of non-epistemic values on scientific justification in communication is immediately clear. But I want to highlight two less obvious implications of this case. First, I agree with John that the VAI is "an appealing ideal," and I do not wish or need to settle whether it is all-things-considered plausible here. However, Mammography Wars raises a curious possibility: if, as the VAI recommends, the values assumed in justification are to align with the values held by putative audiences, then it is far from clear that the NIH Advisory Board reached a problematic conclusion. For it was precisely by considering the values of the audience that the Panel justified their recommendation for screening, which implicitly endorses the claim that roughly saving one life for 1000 women screened for 10 years at the costs of over 250 women harmed constitutes "effective" screening. Yet, I submit, this strains a reasonable understanding of what a plausible benefit-harm ratio for screening would be. Thus, in certain scenarios, such as when the putative audience holds the "incorrect" values, the VAI seems to misfire, justifying scientific claims to be communicated that are misleading at best and harmful at worst.

This is not necessarily a problem for John's account; he formulates the VAI as an "attractive" ideal and not an "obligatory" ideal to follow. Nonetheless, what the considerations above show is that the VAI rests on an assumption that the non-epistemic values of the intended audience are *themselves* apt for the purposes of scientific justification. In many cases, this assumption may be innocuous. As John notes, a physician deciding which side-effects of a drug to report to a patient should appeal to the recipient's values to adjudicate what to communicate. But other times, it will be more complex.

Second, I want to suggest one policy reason to favour being very confident in the claim that screening is effective before asserting it—that is, in demanding a high evidentiary threshold before acting on a claim such as "screening is effective." Contrast The Mammography Wars with another

case-study: infant neuroblastoma screening. In 1991, the American Cancer Society recommended no screening but more research (Murphy *et al.* 1991). This was accepted without a public relations disaster. Why? One major difference was that, when the respective recommendations were published, neuroblastoma screening was not a widespread practice, while breast cancer screening for women in their forties was. So, the latter was viewed as denying a previously accepted practice.

This implies an uncomfortable truth: screening policies are often, in effect, irreversible. Once the public gets used to something, they will not be pleased to have it withdrawn, even if withdrawing it is a better alternative than current practice. As a colleague at a recent conference put it, "Once you've started a screening programme, you are basically f*cked. You're not going to get yourself out of it."[41] The stakeholders underwriting the programme—from patient groups, diagnostic industries, the resources invested by the healthcare system, the health professionals making a living off the practice—are simply in too deep. So before implementing a new programme, we had better be sure we are getting it right. This relates to John's point that we may "need to adopt a complex account of value-aptness, focused on *practices*, rather than *instances* of justification" (7). I agree. My point here is that one way in which the practices of justification in the case of screening must account for institutional inertia and the difficulties inherent in trying to change direction once the ball gets rolling.

5.6 Conclusion

In this chapter I have argued that there is an appropriate role for non-epistemic values in the justification of claims about screening. I discussed the argument from inductive risk, and suggested that some past inferences about screening were problematic because they indirectly invoked contentious values. Further, I addressed an objection from Betz (2013) that science can avoid being value-laden with the use of hedged claims. While intuitively appealing, this objection overlooks some pragmatic nuances related to the communication of scientific claims. I then turned my attention to a recent proposal by John (2018), which argues that the values invoked in scientific justification should be compatible with the values of the intended audience. I raised a worry that John's proposal may go awry when the values of the putative audience are, themselves,

[41] See: Christian Munthe, at minute 13. https://sms.cam.ac.uk/media/2728717

problematic, and I suggested that the strategy could be extended by indexing communication to the appropriate speech act. Finally, I concluded by suggesting a reason to want a high evidentiary threshold before implementing a new screening programme, namely, that once a programme is implemented, it is almost impossible to get rid of.

CONCLUSION

This thesis has examined the conceptual, epistemic, and ethical dimensions of screening. Here is a recap of the central claims, from each chapter. In Chapter 1, "Screening," I set the stage for the thesis by introducing the key concepts central to understanding the screening debate, explaining the logic underlying screening, and motivating why screening is such a topic of fierce controversy. In Chapter 2, "Non-Maleficence," I argued for a novel interpretation of the principle of non-maleficence, *ex ante* DNH, which claims that any screening programme that lowers the *ex ante* interests of some is *prima facie* impermissible. In Chapter 3, "Effectiveness," I discussed the implications of these arguments for the notion of "screening effectiveness." I argued, in particular, that the concept of effectiveness is far more value-laden than commonly presumed, since different ways of thinking about screening effectiveness operationalize different normative principles. In Chapter 4, "Uncertainty," the arguments developed in the previous chapters were extended to contexts of uncertainty, where decision makers are not in a position to assign precise probabilities to outcomes. I explored the implications of this uncertainty for a theory of prospects, the principle of non-maleficence, and a concept of effectiveness. Finally, in Chapter 5, "Underdetermination," I examined the role of non-epistemic values in addressing situations of evidential underdetermination, like screening. I argued that if we focus on the communication of screening claims, then there is a proper role for non-epistemic values when determining the evidentiary thresholds appropriate for the implementation of a new screening programme.

Stepping back slightly, a few central themes have emerged. I will introduce them by way of commenting on the choice of title, which we are now in a better position to understand. The significance of "The Limits of Screening" is two-fold. First, there are important *moral* constraints on screening policymaking that need to be acknowledged. Take arguments for better screening risk-stratification in light of individual variation in cancer risk. Typically, these arguments appeal to cost-effectiveness considerations. The aim on this way of thinking is to maximize QALYs per pound spent (Pashayan *et al.* 2018). But this approach implies a certain moral picture of screening. The picture is one in which screening involves the distribution of benefits, like NICE deciding which drugs to fund given a limited healthcare budget. It is a picture in which the primary harms

of screening are opportunity costs, like NICE deciding to fund a particular cancer drug instead of a heart disease drug.

But there is a more fundamental ethical concern here. Because screening programmes invariably cause harm through false-positives, overdiagnosis, overtreatment, there are real harms involving real people that go beyond mere opportunity costs. So in order to avoid breaching non-maleficence, I argued that screening policymaking should be guided by the *ex ante* DNH principle—screening programmes require a higher justificatory standard than simply being in the interests of some "average" individual; they must be in the interests of *each* individual offered screening. *Ex ante* DNH generates a distinct and distinctively moral argument for better risk-stratification that does not operate via concerns about cost-effectiveness. What is more, *ex ante* DNH points to a more demanding ethical constraint than commonly presumed for the implementation of screening. Screening programmes that lower some individuals' *ex ante* interests may not just be ethically subpar, like an effective altruist donating to the charity that would bring about the seventh-best state of affairs. They may even be ethically *impermissible*, like an effective altruist stealing money from an art charity to buy mosquito nets that save more lives. In this respect, this thesis has endeavoured to articulate some ethical considerations that limit the overzealous implementation of screening.

The second "limit" of screening builds on this last point. One issue this thesis has indirectly gestured at is whether early detection is truly the best strategy to improve cancer care. As the NHS Long-Term Plan exemplified, screening will be the principal strategy in the coming years to tackle cancer. Yet, as we saw, implementing screening is a complex task, and evidence to justify that it is effective is hard to come by. There is a very serious risk of doing more harm than good, as the history of screening illustrates (Raffle and Gray 2007). So, while early detection can surely save some lives from cancer, and while this this is unequivocally a good thing, we need to be attentive to what cost this comes at—focusing solely on the benefits of screening obscures the bigger picture. And the bigger picture is not always rosy. For example, even if there is actually an impact of cancer screening on mortality, this reduction is not so large as to be detected in overall mortality outcome measures.

So much the worse for overall mortality outcome measures? That would be too quick. Sceptics of screening are right to point out that overall mortality is the best outcome measure to track screening effectiveness, because it captures any potential harms of screening. So much the worse for screening? That would also be too quick. The point here is not to deny the importance or promise of early detection; it is merely to caution against a certain form of unwarranted faith in early detection. As Chapter 4 argued, for example, there is a good deal of uncertainty around the rates of overdiagnosis for individuals in a screening programme. So, in making the claim "screening is effective," there is an important sense in which we are making a bet about what the actual chance of being overdiagnosed is for you. It would be unwise to be overly confident in our bets.

There is, of course, much more to be said on philosophical issues in screening. In closing, let me comment on two avenues for future work. In Chapter 2, I argued that the non-maleficence principle has been underappreciated in the context of screening. And in the course of that argument, I expressed scepticism with respect to the normative power of "informed consent" in screening. Communicating about risk, let alone cancer risk, is a notoriously challenging endeavour. I suggested that, with respect to the harms of screening, appealing to consent is rather odd—a bit like asking for your consent to a boxing match in which I am certain I will seriously injure you. You may agree to the match, you may be confident that the match will be fun or challenging in a good way, but this does not mean that my instigating the match does not harm you.

Nor does this mean, however, that the match is entirely impermissible. To claim that acquiring informed consent in screening is difficult may be like the perfect being the enemy of the good. And on one way of thinking, so long as autonomous adults provide genuinely informed consent, we should allow them to participate in the boxing match. We should respect their judgement and decisions as autonomous agents. This reasoning may generalize to screening. But just how much normative weight can we put on this appeal to informed consent? Thinking through this question immediately raises issues. Here is one: what constitutes "genuinely informed consent" in screening? Many experts in risk communication claim that there is no neutral way to present risk (Speigelhalter 2017). Ordering the benefits before the harms or *vice versa*, framing the statistics in a positive or negative lens all can potentially change whether consent is provided (Chwang 2015). So, a substantive task for future work is developing an account of when consent is genuinely

informed in the context of cancer risk communication. This, in turn, requires thinking carefully about how informed consent relates to issues of autonomy in screening, and how autonomy should be valued more generally in a screening context.

I suspect this is a far harder task than most think. In 2018, when I gave a public talk on "The Ethics of Risk Communication," I discussed these issues with an artist tasked with illustrating the central ideas through a different medium—in this case a painting. Over coffee, I expressed some worries I had about the impact of comparative risk information on risk perception. In brief: if I offer you a preventative pill that will reduce your risk of cancer at the cost of some cumbersome side-effects, you should decide whether to take the pill by weighing up the associated benefits and harms. But if I offer you the same pill with the same associated benefits and harms, and tell you in addition that your risk of cancer is *above average*, this will likely change how you perceive your risk of cancer. You are now more likely to accept the pill (Fagerlin *et al*. 2007).

My worry was that this is odd, because the benefits and harms of the preventative pill have not changed. Rationally speaking, all that is relevant to your decision is the associated benefits and harms of the pills for you. Why should whether you have a higher or lower risk of cancer than your neighbour alter your evaluation of whether to take the pill? But the artist was not convinced. Why *not* allow comparative risk information to influence your decision? Isn't comparative information a useful metric to contextualize the benefits and harms of the pill? The artist pressed me further: "Look, for me, whether or not I am above or below average risk determines what it is reasonable to do in the situation." When I explained my worry that her response is simply irrational, the artist let out a sigh of frustration and gave me an annoyed sort of look. And the look stuck with me. It was not the sort of look you give your concerned but mildly overbearing parents when they insist you eat more vegetables. It was the sort of look that you give a well-intentioned friend, who has no experience of rock climbing, when he tries to offer you advice about how to rock climb.

The moral here is a tricky one, because it goes directly to the heart of the ethics of risk communication. On the one hand, there are good ethical reasons to honour the judgment of autonomous agents. Include the comparative risk information, let the recipient decide, and treat her decision with significant respect. Even if her actual choice does not maximize her expected

utility, so what? People make irrational decisions all the time, as the field of behavioural economics has demonstrated (Ariely 2008). On the other hand, there are good ethical reasons to want people to choose in accordance with their best interests. Put the salad before the French fries, keep the vegetables at eye-level, relegate the junk food to foot-level (Thaler and Sunstein 2009). The construction of this choice architecture, for example, is intended to "nudge" people toward choosing what is best for them. But "nudges" can sometimes morph into "shoves." It may be that the inclusion of comparative risk information heavily predisposes individuals to choose a certain option. And this may be a useful tool to promote better health outcomes. Consider, for example, communicating risk in a certain way to increase the uptake of an effective screening programme. The pressing task is finding some middle ground between these two poles—between carving out space for autonomy and formulating policy in specific ways to promote the best outcomes.

A second set of issues for future work concerns the intersection of screening policy and political philosophy. There are some curious links here. We know that screening uptake is lower in more deprived socioeconomic groups (Douglas *et al.* 2016). This seems *prima facie* problematic, because if screening is effective then it is widening health inequalities. However, there is also evidence showing that, at least in the United States, women living in the highest quintile of socioeconomic status had *twice* the rate of breast cancer diagnosis as women in the lowest quintile, even after controlling for the classic breast cancer risk factors (Welch and Brawley 2018). This, too, is *prima facie* problematic. Notice, however, that there is a peculiar upshot here: if Baum (2013) is right that mammography screening leads to more deaths than lives saved, then screening seems to be a vehicle to *minimize* health inequalities.

Of course, nobody justifies screening programmes as a tool to rectify the injustice of health inequalities. And few think that "levelling down" is an acceptable way to achieve egalitarian aims. But the point here is that the relationship between screening and justice is a promising avenue for future work. For example, there is a clear sense in which individuals who, after careful reflection, decline their screening invitation are "responsible" for their actions. On one theory of health justice, though, we should distribute goods in a "responsibility-sensitive" manner (Segall 2010). Health inequalities resulting from people deciding not to brush their teeth are not unjust, on this view, because these people are in some sense "responsible" for the health inequality that arises.

But if people are responsible for their decision to accept or decline a screening offer, to what extent are health inequalities resulting from screening unjust?

These concluding remarks point to the need for more philosophical analyses of screening—or more generally, to a more fully fleshed out philosophy of public health. Nonetheless, I hope this thesis has identified some of the key issues underlying an ethics of screening and provided a useful way to approach thinking about them.

APPENDIX 1

Measurement Challenges

This appendix explains some common challenges to measuring the "effectiveness" of screening. You might think that measuring the benefits and harms of screening is straightforward. You might think, for instance, that it is no more difficult than measuring the benefits and harms of drugs for heart disease. Conduct the meta-analyses and randomized trials, consider the cohort studies and case control studies, and so forth. But screening is different from therapeutic treatments. Screening affects healthy individuals, and when dealing with healthy individuals there are some well-known biases that arise, particularly with measuring the benefits of screening. These biases all supervene on a basic worry: if all you do is measure health in screened people, then regardless of whether screening makes any relevant difference to length or quality of life, the conclusion reached will be one skewed heavily in favour of screening. Consider the following:

> Health Screenee Effect
> John is a wealthy, well-educated, physically active, non-smoking vegetable-eating aficionado. Mike is a low-income, physically inactive, social-smoking factory worker. Research shows that people like John are more likely to accept the screening offer. People like Mike, less so. Researchers measure the health outcomes from screening, and find that screened individuals die less from cancer.

Does cancer screening improve length or quality of life? Without a proper control group, it is difficult to say, because it is difficult to determine whether we would expect better health outcomes in the screened individuals anyway, simply by virtue of the observation that people who turn up for screening tend to be healthier than those who do not (Raffle and Gray 2007).

Length Time Bias

John and Mike both have bowel cancer. John's cancer is relatively indolent, with a good prognosis, and it is picked up by screening. Mike's cancer is aggressive, with a poor prognosis, and was not picked up by screening—it developed rapidly in between screening appointments. Researchers measure the health outcomes from screening, and find that screened individuals die less from cancer.

Does cancer screening improve length or quality of life? Without a proper control group, it is difficult to say, because screening will preferentially detect slow-developing cancers with good prognoses, while failing to detect the aggressive cancers that tend to show up *between* screenings.

Overdiagnosis Bias

John and Mike both have indolent bowel cancer. John attends his screening appointment, Mike skips. Screening picks up a bowel cancer in John but, unbeknownst to him, it is an instance of overdiagnosis. Had John skipped screening like Mike, then his bowel cancer would never have been detected, because John would have died from other causes before the cancer led to symptoms. Researchers measure the health outcomes from screening, and find that screened individuals live longer.

Does cancer screening improve length or quality of life? Without a proper control group, it is difficult to say, because instances of overdiagnosis will increase survival time even though screening did not delay death. For example, this occurred in prostate cancer screening: prostate-specific antigen (PSA) testing picked up many cases of inconsequential or slow-growing disease with low risk of harm. Many of these were overdiagnosed cancers. Survival statistics for PSA testing were therefore good, but this outcome was misleading.

Lead Time Bias

John and Mike both have bowel cancer with the same risk profile. John undergoes screening and is diagnosed at the age of 55. Mike refrains from screening and is diagnosed after symptoms surface at the age of 60. John and Mike, sadly, both pass away from bowel cancer at age 65. Researchers measure the health outcomes from screening, and find that screened individuals live longer.

Does cancer screening improve length or quality of life? Hardly. Screening appears to increase survival time, but this is simply by virtue of starting the clock sooner, by shifting the detection

earlier with no difference in clinical outcomes. Both John and Mike still pass away at the same age. But John was aware of his disease for longer. It seems plausible to think that this causes him a great deal of anxiety, distress, and decreased quality of life. It seems plausible to think that, in the scenario above, John actually ends up worse off than Mike due to screening.

Some of the biases above are easy to address. To avoid the healthy screenee bias, we should properly randomize subjects in clinical studies. To avoid length time bias, we should have proper control groups. To avoid lead time bias, we should not use *only* 5-year survival rates as an outcome for measuring screening effectiveness. So, for example, Cho *et al.* (2014) point out that survival rates must be interpreted in the context of age-adjusted incidence and cancer mortality data. Their strategy to circumvent biased survival rates was to contextualize the outcome against the backdrop of cancer burden, which is comprised of two separate measures. *Incidence*: how many people are diagnosed with cancer, and *mortality*: how many people die from cancer. Using U.S. data from 1975 to 2010, three differing levels of early detection success can then be distinguished:

> Good Progress
> *Increase in Survival, Decrease in Disease Burden*: this is indicative of early detection success. For example, the five-year survival for colorectal cancer patients increased from 48% to 68%, and the burden of cancer dropped: incidence from 60 to 41 per 100,000 and mortality from 28 to 16 per 100,000.
>
> *Little or No Increase in Survival, Decrease in Disease Burden*: sometimes early detection is successful even when survival rates largely remain unchanged. For example, survival was unchanged in cervical cancer, but disease burden fell substantially because screening prevents the incidence of cervical cancer.
>
> Mixed Progress
> *Increase in Survival, Increase in Incidence, Decrease in Mortality*: sometimes survival improves but there are mixed changes in disease burden. For example, 5-year breast cancer survival rates increased from 74% to 91%, and those for prostate cancer increased from 67% to 100%. For both cancers, mortality decreased from 31 to 22 per 100,000. But incidence increased substantially. For breast cancer, from 111 to 162 per 100,000 and for prostate cancer, from 94 to 149 per 100,000. Increased incidence is explained by the introduction of screening programmes, and decreased mortality is explained, in large part, by treatment improvements.

No Progress

Increase in Survival, Increase in Disease Burden: between 1975 and 2010, survival increased for thyroid cancer in women. But there was also an increase in disease burden: mortality did not change, and incidence increased a great deal (from 6.5 to 21 per 100,000). This is explained by more overdiagnosis stemming from ineffective early detection.

Little or No Increase in Survival, Higher Disease Burden: between 1975 and 2010, there was little change in overall survival for lung cancer in women. But there was a sharp increase in disease burden: incidence increased from 25 to 50 per 100,000 and mortality increased from 18 to 38 death per 100,000. These trends are explained by the delayed uptake of cigarette smoking among women compared with men.

APPENDIX 2

Base Rates and Screening

This appendix illustrates how the underlying prevalence of disease can impact the rate of false-positive results. Consider two hypothetical populations of 10,000 individuals. Suppose that, in the first population, the prevalence of cervical cancer is 5% and that in the second, the prevalence is .05%. Moreover, suppose that we screen both populations with a Pap smear with test sensitivity of 95% and test specificity of 95%. The "accuracy" of the screening test will vary widely—even though the test's sensitivity and specificity is equivalent in both cases. This is because of the underlying difference in disease prevalence.

In the first population with 5% prevalence (Table 1), 500 individuals have cervical cancer. With a test sensitivity of 95%, 475 of these 500 will be accurately screened as positive, and 25 will be falsely screened as negative. Amongst the 9500 individuals who do not have cervical cancer, with a 95% test specificity, 9,025 will be accurately screened as negative, and 475 will be falsely screened as positive. This, however, yields an intriguing observation: half of those with positive pap smears do not actually have cervical cancer (calculated in the first column of Table 1 by dividing 475 by 950 total individuals). Although a 50% false-positive rate seems quite high, the problem becomes even more dire in the second population with an even lower cervical cancer prevalence.

	No. of Positive Pap Smears	No. with Negative Pap Smear	Totals
Cervical Cancer	475	25	500
No Cervical Cancer	475	9,025	9,500

Table 1. 5% Prevalence of Cervical Cancer

In the second population with .05% prevalence (Table 2), only 5 individuals have cervical cancer. With a test sensitivity of 95%, it would be reasonable to assume that all five of these individuals would be accurately screened as positive. Amongst the 9,995 individuals who do not have cervical cancer, with a 95% test specificity, 9,495 will be accurately screened as negative, but 500 individuals will be falsely screened as positive. The implications of this are striking: of the 505 positive pap smears, *only 5 will accurately track the presence of cervical cancer*—a 99% false-positive rate.

	No. of Positive Pap Smears	No. with Negative Pap Smear	Totals
Cervical Cancer	5	0	5
No Cervical Cancer	500	9,495	9,995

Table 2. .05% Prevalence of Cervical Cancer

Thus, even if a screening test is highly sensitive and specific, the overall accuracy may be vastly diminished so long as the disease is not common in the population.

BIBLIOGRAPHY

Aktipis, Athena, and Randolph M Nesse. 2013. "Evolutionary Foundations for Cancer Biology." *Evolutionary Applications* 6 (1): 144–59.

Al-Najjar, Nabil and Weinstein, Jonathan. 2009. "The Ambiguity Aversion Literature: A Critical Assessment" *Economics and Philosophy* 25: 249-84.

Alexandrova, Anna. 2017. *A Philosophy for the Science of Well-Being.* Oxford: Oxford University Press.

Alexandrova, Anna. 2018. "Can the Science of Well-Being Be Objective?" *British Journal for the Philosophy of Science* 69(2): 421-445.

Andersen, Holly. 2012. "Mechanisms: What are they evidence for in evidence-based medicine?" *Journal of Evaluation in Clinical Practice* 18: 992-999.

Ariely, Dan. 2008. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. HarperCollins.

Aronowitz. Robert. 2015. *Risky Medicine.* Chicago: University of Chicago Press.

Ashcroft, Richard. 2002. "What is clinical effectiveness?" *Studies in History and Philosophy of the Biological and Biomedical Sciences* 33(2): 219e233.

Attema, Arthur, Bleichrodt, Han, and L'Haridon, Olivier. 2018. "Ambiguity preferences for health." *Health Economics* 27(11): 1699–1716.

Autier, P, M Boniol, A Gavin, and L J Vatten. 2011. "Breast Cancer Mortality in Neighbouring European Countries with Different Levels of Screening but Similar Access to Treatment: Trend Analysis of WHO Mortality Database." *BMJ* 343: d4411.

Baillon, Aurélian and Bleichrodt, Han. 2015. “Testing Ambiguity Models through the Measurement of Probabilities for Gains and Losses.” *American Economic Journal: Microeconomics* 7(2): 77-100.

Barros, D Benjamin. 2008. “Natural Selection as a Mechanism.” *Philosophy of Science* 75 (3): 306–22.

Baum, Michael. 2013. “Harms From Breast Cancer Screening Outweigh Benefits if Death Caused by Treatment Is Included.” *BMJ* 346: f385.

Beatty, John. 1995. “The evolutionary contingency thesis.” In: Wolters G, Lennox JG (eds) *Concepts, theories, and rationality in the biological sciences*. University of Pittsburgh Press, Pittsburgh, pp 45–81.

Beatty, John. 2006. “Replying life’s tape.” *Journal of Philosophy* 103(7): 336-362.

Beauchamp, Tom, & Childress, James. 2013. *Principles of biomedical ethics*. 7th Ed. Oxford: Oxford University Press.

Bertolaso, Marta. 2016. *Philosophy of Cancer: A Dynamic and Relational View*. Springer.

Betz, Gregor. 2013. “In Defence of the Value Free Ideal.” *European Journal for Philosophy of Science* 3 (2): 207–20.

Biddle, Justin. 2013. “State of the field: Transient underdetermination and values in science.” *Studies in History and Philosophy of Science* 44:124-133.

Biddle, Justin. 2016. “Inductive Risk, Epistemic Risk, and Overdiagnosis of Disease.” *Perspectives on Science* 24(2): 192–205.

Bissell, Mina J, and William C Hines. 2011. “Why Don't We Get More Cancer? a Proposed Role of the Microenvironment in Restraining Cancer Progression.” *Nature Reviews Cancer* 17

(3): 320–29.

Bleyer, Archie, and H Gilbert Welch. 2012. “Effect of Three Decades of Screening Mammography on Breast-Cancer Incidence.” *New England Journal of Medicine* 367 (21): 1998–2005.

Boorse, Christopher. 1977. “Health as a theoretical concept.” *Philosophy of Science* 44(4): 542-573.

Bradley, Richard. 2017. *Decision Theory with a Human Face.* Cambridge: Cambridge University Press.

Broadbent, Alex. 2013. *Philosophy of epidemiology*. Basingstoke: Palgrave Macmillan.

Brodersen, John, and V D Siersma. 2013. “Long-Term Psychosocial Consequences of False-Positive Screening Mammography.” *The Annals of Family Medicine* 11 (2): 106–15.

Broome, John. 1998. “Kamm on Fairness.” *Philosophy and Phenomenological Research* 58: 955–61.

Burki, Talha. 2018. “The Cochrane board votes to expel Peter Gøtzsche.” *The Lancet* 392(10153): 1103-4.

Cairns, John. 1975. “Mutation Selection and the Natural History of Cancer.” *Nature* 255: 1–4.

Camerer, Colin and Weber, Martin. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty* 5(4): 325‑370.

Campbell, P. J., E. D. Pleasance, P. J. Stephens, E. Dicks, R. Rance, I. Goodhead, G. A. Follows et al. 2008. “Subclonal phylogenetic structures in cancer revealed by ultra-deep

sequencing." *Proceedings of the National Academy of Sciences of the United States of America* 105:13081–13086.

Cartwright, Nancy. 2007. "Are RCTs the gold standard*?" BioSocieties* 2: 11-20.

Cartwright, Nancy. 2009. "What Are Randomised Controlled Trials Good for?" *Philosophical Studies* 147 (1): 59–70.

Cartwright, Nancy. 2011. "Predicting 'It will work for us': (Way) beyond statistics." In F. R. Phyllis McKay Illari, & Jon Williamson (Eds.), *Causality in the sciences*. Oxford Scholarship Online.

Cartwright, Nancy. 2012. "Will this policy work for you? Predicting effectiveness better: How philosophy helps." *Philosophy of Science* 79: 973-989.

Castle, Philip, Schiffman, Mark, Wheeler, Cosette, Solomon, Diane. 2009. "Evidence for frequent regression of cervical intraepithelial neoplasia-grade 2." *Obstetrics and Gynecology* 113(1): 18–25.

Catalona, William. 1993. "Screening for prostate cancer: Enthusiasm." *Urology* 42(2):113–115.

Chalmers, David. 2011. "Verbal Disputes." *Philosophical Review* 120(4): 515-566.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. New York: Oxford University Press.

Childe, Charles. 1907. *The Control of a Scourge, or How Cancer is Curable*. New York: E.P. Dutton & Company.

Cho, Hyunsoon, Mariotto, Angela, Schwartz, Lisa, Luo, Jun, and Woloshin, Steven. 2014. "When Do Changes in Cancer Survival Mean Progress? the Insight From Population Incidence and Mortality." *JNCI Monographs* 2014 (49): 187–97.

Chwang, Eric. 2015. “Consent's Been Framed: When Framing Effects Invalidate Consent and How to Validate It Again.” *Journal of Applied Philosophy* 33 (3): 270–85.

Cintolo-Gonzalez, Jessica, Danielle Braun, Amanda Blackford, Emanuele Mazzola, *et al.* 2015. “Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications.” *Breast Cancer Research and Treatment* 164(2): 263-284.

Clarke, Brendan, Donald Gillies, Phyllis Illari, Federica Russo, and Jon Williamson. 2014. “Mechanisms and the Evidence Hierarchy.” *Topoi* 33 (2): 339–60.

Cooper, Rachel. 2002. “Disease.” *Studies in History and Philosophy of the Biological and Biomedical Sciences* 33: 263-282.

Croswell, Jennifer M, David F Ransohoff, and Barnett S Kramer. 2010. “Principles of Cancer Screening: Lessons From History and Study Design Issues.” *Seminars in Oncology* 37 (3): 202–15.

Damiano R, Lorenzo GD, Cantiello F*, et al*. 2007. “Clinicopathologic features of prostate adenocarcinoma incidentally discovered at the time of radical cystectomy: an evidence-based analysis.” *European Urology* 52(3): 648–657.

Daniels, Norman. 2015. “Can There be Moral Force to Favoring an Identified over a Statistical Life?” in *Identified versus Statistical Lives: An Interdisciplinary Perspective* (eds.) Glenn Cohen, Norman Daniels, and Nir Eyal. Oxford: Oxford University Press.

Darby, Sarah, Ewertz, Marianne, McGale, Paul, *et al.* 2013. “Risk of Ischemic Heart Disease in Women after Radiotherapy for Breast Cancer.” *New England Journal of Medicine* 368: 987-998.

Deaton, Angus, and Nancy Cartwright. 2018. “Understanding and Misunderstanding Randomized Controlled Trials.” *Social Science & Medicine* 210: 2–21.

Diener, E., R. Lucas, U. Schimmack and J. Helliwell. 2008. *Well-being for Public Policy*. New York: Oxford University Press.

Dobell, Horace. 1861. "Lectures on the Germs and Vestiges of Disease, and on the Prevention of the Invasion and Fatality of Disease by Periodical Examinations." London: Churchill.

Douglas, Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67(4): 559–579.

Douglas, Heather. 2009. *Science, Policy and the Value-Free Ideal.* Pittsburgh: University of Pittsburgh Press.

Douglas, Elaine, Waller, Jo, Duffy, Stephen, and Wardle, Jane. 2016. "Socioeconomic inequalities in breast and cervical screening coverage in England: are we closing the gap?" *Journal of Medical Screening* 23(2): 98-103.

Elga, Adam. 2010. "Subjective Probabilities Should Be Sharp." *Philosophers Imprint* 10 (5): 1–11.

Ellsberg, Daniel. 1961. "Risk, Ambiguity, and the Savage Axioms." *Quarterly Journal of Economics* 75: 643–669.

Emerson, Haven. 1923. "Periodic medical examinations of apparently healthy persons." *JAMA* 80:1376– 1381.

Ereshefsky, Marc. 2009. "Defining 'health' and 'disease'." *Studies in History and Philosophy of the Biological and Biomedical Sciences* 40: 221-227.

Esserman, Laura J, Ian M Thompson, and Brian Reid. 2013. "Overdiagnosis and Overtreatment in Cancer." *JAMA* 310 (8): 797–2.

Esserman, Laura, Shieh, Yiwey, & Thompson, Ian. 2010. "Rethinking Screening for Breast Cancer and Prostate Cancer." *Journal of the American Medical Association* 302(15):

1685–1692.

Fagerlin, Angela, Zikmund-Fisher, Brian, Ubel, Peter. 2007. “If I’m better than average, then I’m ok?”: Comparative information influences beliefs about risk and benefits.” *Patient Education and Counselling* 69: 140-44.

Feinberg, Joel. 1986. *Harm to Others*. Oxford: Oxford University Press.

Foot, Phillipa. 1980. The problem of abortion and the doctrine of double effect. Reprinted in: Steinbock B,ed. Killing and letting die. Englewood Cliffs: Prentice-Hall: 156-65.

Forbes, Lindsay JL, Amanda-Jane Ramirez, the Expert group on Information about Breast Screening *et al*. 2014. “Offering Informed Choice About Breast Screening.” *Journal of Medical Screening* 21 (4): 194–200.

Franco, Paul. 2017. “Assertion, Nonepistemic Values, and Scientific Practice.” *Philosophy of Science* 84: 160-180.

Frick, Johann. 2013. “Treatment versus Prevention in the Fight against HIV/AIDS and the Problem of Identified versus Statistical Lives.” in *Identified versus Statistical Lives: An Interdisciplinary Perspective* (eds.) Glenn Cohen, Norman Daniels, and Nir Eyal. Oxford: Oxford University Press.

Frick, Johann. 2015. “Contractualism and Social Risk.” *Philosophy and Public Affairs* 43 (3): 175–223.

Fuller, Jonathan. 2018. “Meta-Research Evidence for Evaluating Therapies.” *Philosophy of Science* 85: 767–80.

Fuller, Jonathan, and Luis J Flores. 2015. “The Risk GP Model: the Standard Model of Prediction in Medicine.” *Studies in History and Philosophy of Biol & Biomed Sci* 54 (C): 49–61.

Gigerenzer, Gerd. 2014. “Breast Cancer Screening Pamphlets Mislead Women.” *BMJ* 348: g2636.

Gigerenzer, Gerd. 2015. “Towards a paradigm shift in cancer screening: informed citizens instead of greater participation.” *BMJ* 350: h2175–h2175.

Gilboa, Itzhak, Postlewaite, Andrew, and Schmeidler, David. 2009. “Is it always rational to satisfy Savage’s axioms?” *Economics & Philosophy* 25: 285–96.

Gillon, Raanan. 1985. “Primum non nocere and the principle of non-maleficence.” *BMJ* 291: 130-131.

Goodin, Robert. 1995. *Utilitarianism as a Public Philosophy*. New York: Cambridge University Press.

Gøtzsche, Peter. 2009. Breast screening: the facts— or maybe not. *BMJ* 338: 446-48.

Gøtzsche, Peter. 2012. *Mammography Screening: Truth, Lies, and Controversy.* Radcliffe Publishing.

Gøtzsche, Peter and Jørgensen, Karsten. 2013. “Screening for breast cancer with mammography.” *Cochrane Database of Systematic Reviews* 6 Art. No.: CD001877. DOI: 10.1002/14651858.CD001877.pub5.

Gould, George. 1900. “A system of personal biologic examinations the condition of adequate medical and scientific conduct of life.” *JAMA* 284: 134–138.

Gould, Stephen J. 1989. *Wonderful life: the Burgess Shale and the nature of history.* Norton, New York.

Grice, H. P. 1957. “Meaning.” *Philosophical Review* 66 (3): 377–88.

Hájek, Alan. 2007. “The Reference Class Problem is Your Problem Too.” *Synthese* 156(3): 563-585.

Halstead, John. 2016. “The Numbers Always Count.” *Ethics* 126: 789-802.

Han, Paul. 1997. “Historical changes in the objectives of the periodic health examination.” *Ann Intern Med* 127(10): 910–917.

Hanahan, Douglas, and Robert A Weinberg. 2011. “Hallmarks of Cancer: the Next Generation.” *Cell* 144 (5): 646–74.

Hanley, James. 2011. “Measuring Mortality Reductions in Cancer Screening Trials.” *Epidemiologic Reviews* 33: 36-45.

Hanna, Jason. 2011. “Consent and the Problem of Framing Effects.” *Ethical Theory and Moral Practice* 14 (5): 517–31.

Hardisty, David and Weber, Elke. 2009. “Discounting future green: Money versus the environment.” *Journal of Experimental Psychology: General* 138(3): 329-340.

Hare, Caspar. 2010. “Take the Sugar.” *Analysis* 70 (2): 237–47.

Hare, Caspar. 2013. *The Limits of Kindness*. Oxford: Oxford University Press.

Hare, Caspar. 2017. “Risk and Radical Uncertainty in HIV Research.” *Journal of Medical Ethics* 43 (2): 87–89.

Harris, Russell, George Sawaya, Virginia Moyer, and Ned Calonge. 2011. “Reconsidering the Criteria for Evaluating Proposed Screening Programs: Reflections From 4 Current and Former Members of the U.S. Preventive Service Task Force.” *Epidemiologic Reviews* 33: 20–35.

Harsanyi, John. 1955. “Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility.” *Journal of Political Economy* 63: 309–21.

Hausman, Dan. 2015. *Valuing Health: Well-Being, Freedom, and Suffering*. Oxford: Oxford University Press.

Hawkes, Nigel. 2018. "Breast Cancer Screening Error: Fatal Mistake or Lucky Escape?." *BMJ* 361: k2036–2.

Heal, Geoffrey and Millner, Antony. 2014. "Uncertainty and Decision Making in Climate Change Economics." *Review of Environmental Economics and Policy* 8(1): 120-137.

Hoffmann, Tammy and Del Mar, Chris. 2015. "Patients' Expectations of the Benefits and Harms of Treatments, Screening, and Tests." *JAMA Internal Medicine* 175(2): 274-286.

Howick, Jeremy. 2011. *The Philosophy of Evidence-Based Medicine*. Chichester, United Kingdom: Wiley-Blackwell.

Hux, Janet, and C. David Naylor. 1995. "Communicating the Benefits of Chronic Preventive Therapy: Does the Format of Efficacy Data Determine Patients' Acceptance of Treatment?" *Medical Decision Making* 15: 152–57.

Jacobs, I., Menon, U., Ryan, A., Gentry-Maharaj, A., Burnell, M., Kalsi, Jatinderpal., Amso, N. et al. 2016. "Ovarian Cancer Screening and Mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a Randomised Controlled Trial." *The Lancet* 387: 945–56.

Jeffrey, Richard. 1956. "Valuation and Acceptance of Scientific Hypotheses." *Philosophy of Science* 22 (3): 237–46.

Jenni, Karen, and George Loewenstein. 1997. "Explaining the 'Identifable Victim.'" *Journal of Risk and Uncertainty* 14 (3): 235–57.

John, Stephen. 2014. "Inductive risk and the contexts of communication." *Synthese* 192(1): 79–96.

John, Stephen. 2014. "Risk, Contractualism, and Rose's "Prevention Paradox"." *Social Theory*

and Practice 40 (1): 28–50.

John, Stephen. 2018. “Science, truth and dictatorship: Wishful thinking or wishful speaking?” *Studies in History and Philosophy of Science*, 1–9.

Justman, Stewart. 2010. “Uninformed Consent: Mass Screening for Prostate Cancer.” *Bioethics* 26 (3): 143–48.

Juth, Niklas and Munthe, Christian. 2012. *The Ethics of Screening in Healthcare and Medicine*. New York: Springer.

Kahneman, Daniel, A.B. Krueger, D.A. Schkade, N. Schwarz, and A.A. Stone. 2004. “A survey method for characterizing daily life experience: The day reconstruction method.” *Science* 306(5702): 1776-80.

Kaivanto, Kim, and Daniel Steel. 2019. “Adjusting Inferential Thresholds to Reflect Nonepistemic Values.” *Philosophy of Science* 86: 255–85.

Kamm, Frances. 2015. “Cost Effectiveness Analysis and Fairness.” *Journal of Practical Ethics* 3(1): 1-14.

Kavka, George. 1990. “Some Social Benefits of Uncertainty.” *Midwest Studies in Philosophy* XV: 311-326.

Keynes, John Maynard. 1937. “The General Theory of Employment.” *The Quarterly Journal of Economics* 51: 209–23.

Kitcher, Philip. 2001. *Science, truth, and democracy*. New York: Oxford University Press.

Kopans, Daniel, Smith, Robert, and Duffy, Stephen. 2011. “Mammographic Screening and ‘Overdiagnosis.’” *Radiology* 260(3): 616–20.

Laplane, Lucie. 2016. *Cancer Stem Cells: Philosophy and Therapies.* Harvard University Press.

Lerner, Barron. 2001. *The Breast Cancer Wars*. New York: Oxford University Press.

Levi, Isaac. 1960. "Must the Scientist Make Value Judgments?" *Journal of Philosophy* 57 (11): 345–57.

Levi, Isaac. 1962. "On the Seriousness of Mistakes." *Philosophy of Science* 29 (1): 47–65.

Lewens, Tim (Ed.). 2007. *Risk: Philosophical Perspectives*. London: Routledge.

Lewens, Tim. 2019. "The division of advisory labour: the case of "mitochondrial donation." *European Journal of Philosophy of Science* 9: 10–33.

Lipinski, Kamil A, Louise J Barber, Matthew N Davies, Matthew Ashenden, Andrea Sottoriva, and Marco Gerlinger. 2016. "Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine." *TRENDS in CANCER* 2 (1): 49–63.

Longino, Helen. 1990. *Science as Social Knowledge*. Princeton, NJ: Princeton University Press.

Machamer, Peter, Lindley Darden, and Carl Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67 (1): 1–25.

Mahtani, Anna. 2017. "The Ex Ante Pareto Principle." *Journal of Philosophy* 114 (6): 303–23.

Malenka, David, John Baron, Sarah Johansen, Jon Wahrenberger, and Jonathan Ross. 1993. "The Framing Effect of Relative and Absolute Risk." *Journal of General Internal Medicine* 8(10): 543– 48.

Malm, Heidi. 1999. "Medical Screening and the Value of Early Detection: When Unwarranted Faith Leads to Unethical Recommendations." *Hastings Center Report* 1: 26–37.

Marcum, James A. 2005. "Metaphysical Presuppositions and Scientific Practices: Reductionism

and Organicism in Cancer Research." *International Studies in the Philosophy of Science* 19 (1): 31–45.

Marmot, Michael. 2017. "Social justice, epidemiology and health inequalities." *European Journal of Epidemiology* 32(7): 537-546.

Marmot, Michael, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. 2012. "The benefits and harms of breast cancer screening: an independent review." *The Lancet* 380:1778–86.

Marshall, Eliot. 2014. "Dare to Do Less." *Science* 343 (6178): 1454–56.

Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. 2015. "Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin." *Science* 348: 880–86.

McConwell, Alison. 2017. "Contingency and Individuality: a Plurality of Evolutionary Individuality Types." *Philosophy of Science* 84: 1104–16.

McConwell, Alison K, and Adrian Currie. 2016. "Gouldian Arguments and the Sources of Contingency." *Biology & Philosophy* 32 (2): 243–61.

Mitchell, Sandra. 2004. "The Prescribed and Proscribed Values in Science Policy." *In Science, Values, and Objectivity*, ed. Peter Machamer and Georen Wolters, 245–55. Pittsburgh: University of Pittsburgh Press.

Mill, John Stuart. 1869. *On Liberty.* London: Longman, Roberts & Green.

Mooi, W.J., Peeper, D.S. 2006. "Oncogene-induced cell senescence—halting on the road to cancer." *New England Journal of Medicine* 355(10): 1037–1046.

Mukherjee, Siddhartha. 2011. *The Emperor of All Maladies*. Scribner.

Murphy, S.B., Cohn, S.L., Craft, A.W., *et al.* 1991. “Do children benefit from mass screening for neuroblastoma? Consensus Statement from the American Cancer Society Workshop on Neuroblastoma Screening.” *The Lancet* 337: 344-346.

Newman, David H. 2010. “Screening for Breast and Prostate Cancers: Moving Toward Transparency.” *JNCI Journal of the National Cancer Institute* 102 (14): 1008–11.

Nord, Erik. (1999). *Cost–value Analysis in Healthcare*. Cambridge, Cambridge University Press.

Nowell, Peter. 1976. “The Clonal Evolution of Tumor Cell Populations.” *Science* 194 (4260): 23–28.

Nozick, Robert. 1974. *Anarchy, State and Utopia*. New York: Basic Books.

Nyström, L., Rutqvist, L.E., Wall, S., Lindgren, A., Lindqvist, M., Ryden, S., *et al.* 1993. Breast cancer screening with mammography: overview of Swedish randomised trials. *The Lancet* 341(8851): 973–8.

Okasha, Samir. 2006. *Evolution and the Levels of Selection*. Clarendon Press, Oxford.

Orzack, Steven Hecht and Forber, Patrick. 2017. "Adaptationism." The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2017/entries/adaptationism/>.

Otsuka, Michael and Alex Voorhoeve. 2009. “Why It Matters that Some Are Worse Off Than Others: An Argument against the Priority View.” *Philosophy and Public Affairs* 37(2): 171-199.

Ove Hansson, Sven. 2006. “Economic (Ir)rationality in Risk Analysis.” *Economics and Philosophy* 22: 231-241.

Owens, S. 2015. *Knowledge, policy and expertise*. Oxford: Oxford University Press.

Park, S. Y., M. Gonen, H. J. Kim, F. Michor, and K. Polyak. 2010. "Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype." *The Journal of Clinical Investigation* 120: 636–644.

Pashayan, Nora, Steve Morris, Fiona J Gilbert, and Paul D P Pharoah. 2018. "Cost-Effectiveness and Benefit-to-Harm Ratio of Risk-Stratified Screening for Breast Cancer." *JAMA Oncology* 4 (11): 1504–10.

Pellegrino, Edmund. 2001. "The internal morality of clinical medicine: a paradigm for the ethics of the helping and healing professions." *The Journal of Medicine and Philosophy* 26 (6): 559-579.

Penston, James. 2011. "Should we use total mortality rather than cancer specific mortality to judge cancer screening programmes? Yes." *BMJ* 343: d6395.

Philipson, Tomas, Eber, Michael, Lakdawalla, Darius, Corral, Mitra, *et al*. 2012. "An analysis of whether higher health care spending in the United States versus Europe is 'worth it' in the case of cancer." *Health Affairs* 31(4): 667–675.

Plutynski, Anya. 2018. "Safe or Sorry? Cancer Screening and Inductive Risk." In *Exploring Inductive Risk: Case Studies of Values in Science* (eds.) Elliott, Kevin and Richards, Ted. Oxford: Oxford University Press.

Plutynski, Anya. 2018. *Explaining Cancer: Finding Order in Disorder*. Oxford: Oxford University Press.

Powell, Russell. 2009. "Contingency and Convergence in Macroevolution: a Reply to John Beatty." *Journal of Philosophy* 106 (7): 390–403.

Prasad, Vinay, Jeanne Lenzer David H Newman. 2016. "Why Cancer Screening Has Never Been

Shown to ‘Save Lives’—and What We Can Do About It.” *BMJ* 352: h6080.

Puliti, Donella, Stephen W Duffy, Guido Miccinesi, Harry De Koning, Elsebeth Lynge, Marco Zappa, and Eugenio Paci. 2012. “Overdiagnosis in Mammographic Screening for Breast Cancer in Europe: a Literature Review.” *Journal of Medical Screening* 19: 42–56.

Quanstrum, Kerianne, and Rodney Hayward. 2010. “Lessons From the Mammography Wars.” *The New England Journal of Medicine* 363: 1076–79.

Raffle, Angela and Gray, Muir. 2007. *Screening: Evidence and Practice*. Oxford: Oxford University Press.

Rawls, John. 1999. *A Theory of Justice.* Rev. ed. Oxford: Oxford University Press.

Reid, Lynette. 2017. “Truth or Spin? Disease Definition in Cancer Screening.” *Journal of Medicine and Philosophy* 42 (4): 385–404.

Reiser, Stanley. 1978. “The emergence of the concept of screening for disease.” *Milbank Mem Fund Q Health Soc* 56(4): 403–425.

Richardson, Henry. 1990. “Specifying Norms as a Way to Resolve Concrete Ethical Problems.” *Philosophy & Public Affairs* 19(4): 279-310.

Rogers, Wendy, Craig, Wendy, Entwistle, Vicky. 2017. “Ethical issues raised by thyroid cancer overdiagnosis: A matter for public health? *Bioethics* 31(8): 590-598.

Rose, Geoffrey. 2008. *The Strategy of Preventive Medicine*. Oxford: Oxford University Press.

Rowe, Thomas, and Alex Voorhoeve. 2019. “Egalitarianism Under Severe Uncertainty.” *Philosophy and Public Affairs* 46 (3): 239–68.

Rudner, Richard. 1953. “The Scientist qua Scientist Makes Value Judgments.” *Philosophy of Science* 20 (1): 1–6.

Russell, Bertrand. 1912. *The Problems of Philosophy*. Arc Manor, Rockville, MD.

Russo, Federica, and Jon Williamson. 2007. “Interpreting Causality in the Health Sciences.” *International Studies in the Philosophy of Science* 21 (2): 157–70.

Ryan, R.M., & Deci, E.L. 2001. “On happiness and human potentials: A review of research on hedonic and eudaimonic well-being.” *Annual Review of Psychology* 52(1): 141-66.

Sadler, Lynn, Saftlas, Audrey, Wang, Wenquan. *et al.* 2004. “Treatment for cervical intraepithelial neoplasia and risk of preterm delivery.” *JAMA* 291(17): 2100–2106.

Saquib, N, J Saquib, and J P Ioannidis. 2015. “Does Screening for Disease Save Lives in Asymptomatic Adults? Systematic Review of Meta-Analyses and Randomized Trials.” *International Journal of Epidemiology* 44 (1): 264–77.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schwartz, Peter. 2014. “Small Tumors as Risk Factors Not Disease.” *Philosophy of Science* 81: 986–98.

Segall, Shlomi, 2010. *Health, Luck, and Justice*. Princeton: Princeton University Press.

Shieh, Yiwey, Eklund, Martin, Sawaya, George, Black, William, Kramer, Barnett, & Esserman, Laura. 2016. “Population-based screening for cancer: hope and hype.” *Nature Reviews Clinical Oncology* 13(9): 550-565.

Skipper, Robert A., Jr., and Roberta L. Millstein. 2005. “Thinking about Evolutionary Mechanisms: Natural Selection.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 327–347.

Sliwa, Paulina, and Sophie Horowitz. 2015. “Respecting All the Evidence.” *Philosophical Studies* 172 (11): 2835–58.

Slovic, Paul. 1987. “Perception of Risk.” *Science* 236(4799): 280-5.

Slovic, Paul, Finucane, Melissa, Peters, Ellen, Macgregor, Donald. 2004. “Risk as Analysis and Risk as Feelings: Some Thoughts about Affect, Reason, Risk, and Rationality.” *Risk Analysis* 24(2): 311-22.

Smith RA *et al*. 2015. “Cancer Screening in the United States, 2015: A Review of Current American Cancer Society Guidelines and Current Issues in Cancer Screening.” *CA: A Cancer Journal for Clinicians* 65(1): 30-54.

Sober, Elliot. 2009. Absence of evidence and evidence of absence: evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads. *Philosophical Studies* 143: 63-90.

Sorensen, L, D Gyrd-Hansen, IS Kristiansen, J Nexoe, and JB Nielsen. 2008. “Laypersons’ Understanding of Relative Risk Reductions: Randomised Cross-Sectional Study.” *BMC Med- ical Informatics and Decision Making* 8 (31). doi:10.1186/1472-6947-8-31

Speigelhalter, David. 2017. “Risk and Uncertainty Communication.” *Annual Review of Statistics and Its Application* 4: 31-60.

Sprenger, Jan and Stegenga, Jacob. 2017. “Three Arguments for Absolute Outcome Measures.” *Philosophy of Science* 84: 840-852.

Steel, Daniel. 2008. *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.

Steele, Katie. 2012. "The Scientist qua Policy Advisor Makes Value Judgments." *Philosophy of Science* 79: 893–904.

Steele, Robert and Brewster, David. 2011. "Should we use total mortality rather than cancer specific mortality to judge cancer screening programmes? No." *BMJ* 343: d6397.

Stegenga, Jacob. 2015. "Effectiveness of medical interventions." *Studies in History and Philosophy of Biological and Biomedical Sciences* 54(C): 34-44.

Stegenga, Jacob. 2018. *Medical Nihilism*. Oxford, United Kingdom: Oxford University Press.

Stratton, Michael R, Peter J Campbell, and P Andrew Futreal. 2009. "The Cancer Genome." *Nature* 458 (7239): 719–24.

Tannock, Ian, and John Hickman. 2016. "Limits to Personalized Cancer Medicine." *New England Journal of Medicine* 375 (13): 1289–1894.

Taurek, John. 1977. "Should the numbers count?" *Philosophy and Public Affairs* 6(4): 293-316.

Thaler, Richard and Sunstein, Cass. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New York: Penguin Books.

Thompson, Christopher. 2017. "Rose's Prevention Paradox." *Journal of Applied Philosophy* 35(2): 242-256.

Thomson, Judith Jarvis. 1993. "Goodness and Utilitarianism." *Proceedings and Addresses of the American Philosophical Association* 67: 145–59.

Tomlin, Patrick. 2017. “On Limited Aggregation.” *Philosophy and Public Affairs* 45(3): 232-260.

Trautmann , Sefan and van de Kuilen, Gijs. 2015. “Ambiguity Attitudes.” In *The Wiley Blackwell Handbook of Judgment and Decision Making*, ed. Gideon Keren and George Wu. 89-116. Chichester: Wiley.

Tversky, Amos, and Daniel Kahneman. 1981. “The Framing of Decisions and the Psychology of Choice.” *Science* 211 (4481): 453–58.

Usher-Smith, Juliet A, Barbora Silarova, Artitaya Lophatananon, Robbie Duschinsky, Jackie Campbell, Joanne Warcaba, and Kenneth Muir. 2017. “Responses to Provision of Personalised Cancer Risk Information: a Qualitative Interview Study with Members of the Public.” *BMC Public Health* 17: 977–86.

Usher-Smith, Juliet A, Barbora Silarova, Stephen J Sharp, Katie Mills, and Simon J Griffin. 2018. “Effect of Interventions Incorporating Personalised Cancer Risk Information on Intentions and Behaviour: a Systematic Review and Meta-Analysis of Randomised Controlled Trials.” *BMJ Open* 8 (1): e017717–14.

Van den Bruel, Ann, Jones, Caroline, Yang, Yaling, Oke, Jason, Hewitson, Paul. 2015. “People's willingness to accept overdetection in cancer screening: population survey.” *BMJ* 350: h980.

Verweij, Marcel. 1999. “Medicalization as a Moral Problem for Preventive Medicine.” *Bioethics* 13 (2): 89–113.

Voorhoeve, Alex. 2014. “How Should We Aggregate Competing Claims?” *Ethics* 125: 64-87.

Voorhoeve, Alex, and Fleurbaey, Marc. 2016. “Priority or Equality for Possible People?.” *Ethics* 126: 929–54.

Walker, Mary Jean, and Wendy Rogers. 2017. “Defining Disease in the Context of Overdiagnosis.” *Medicine, Health Care and Philosophy* 20 (2): 269–80.

Welch, Gilbert. 2006. *Should I Be Tested for Cancer? Maybe Not and Here’s Why.* Berkeley: University of California Press.

Welch, Gilbert, and Albertsen, Peter. 2009. “Prostate Cancer Diagnosis and Treatment After the Introduction of Prostate-Specific Antigen Screening: 1986-2005.” *JNCI Journal of the National Cancer Institute* 101 (19): 1325–29.

Welch, Gilbert, and Black, William. 2010. “Overdiagnosis in Cancer.” *JNCI Journal of the National Cancer Institute* 102 (9): 605–13.

Welch, Gilbert and Brawley, Otis. 2018. “Scrutiny-Dependent Cancer and Self-fulfilling Risk Factors.” *Annals of Internal Medicine* 168(2): 143-5.

Wilholt, Torsten. 2009. “Bias and values in scientific research.” *Studies in History and Philosophy of Science* 40: 92–101.

Wilholt, Torsten. 2013. “Epistemic trust in science.” *The British Journal for the Philosophy of Science* 64(2): 233–253.

Willig, Carla. 2011. “Cancer Diagnosis as Discursive Capture: Phenomenological Repercussions of Being Positioned Within Dominant Constructions of Cancer.” *Social Science & Medicine* 73 (6): 897–903.

Wilson, James. 2009. “Towards a Normative Framework for Public Health Ethics and Policy.” *Public Health Ethics* 2 (2): 184–194.

Wilson, J. M. and Jungner, Y. G. 1968. “Principles and practice of screening for disease.” *Public Health Papers* 34: 1–163.

Woloshin, Steven and Schwartz, Lisa. 2012. “How a charity oversells mammography.” *BMJ* 345: e5132.

Wong, T Y William. 2019. “The Evolutionary Contingency Thesis and Evolutionary Idiosyncrasies.” *Biology & Philosophy* 34: 22.

Wright, Caroline, and Zimmern, Ron. 2014. “Conceptual Issues for Screening in the Genomic Era - Time for an Update?” *Epidemiology Biostatistics and Public Health* 11(4): e9944.